

Fine-Grained Facial Expression Recognition in Multiple Smiles

Zhijia Jin ¹ , Xiaolu Zhang ^{2,*}, Jie Wang ¹, Xiaolin Xu ¹ and Jiangjian Xiao ²¹ Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China² Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo 315201, China

* Correspondence: zhangxiaolu@nimte.ac.cn

Abstract: Smiling has often been incorrectly interpreted as “happy” in the popular facial expression datasets (AffectNet, RAF-DB, FERPlus). Smiling is the most complex human expression, with positive, neutral, and negative smiles. We focused on fine-grained facial expression recognition (FER) and built a new smiling face dataset, named Facial Expression Emotions. This dataset categorizes smiles into six classes of smiles, containing a total of 11,000 images labeled with corresponding fine-grained facial expression classes. We propose Smile Transformer, a network architecture for FER based on the Swin Transformer, to enhance the local perception capability of the model and improve the accuracy of fine-grained face recognition. Moreover, a convolutional block attention module (CBAM) was designed, to focus on important features of the face image and suppress unnecessary regional responses. For better classification results, an image quality evaluation module was used to assign different labels to images with different qualities. Additionally, a dynamic weight loss function was designed, to assign different learning strategies according to the labels during training, focusing on hard yet recognizable samples and discarding unidentifiable samples, to achieve better recognition. Overall, we focused on (a) creating a novel dataset of smiling facial images from online annotated images, and (b) developing a method for improved FER in smiling images. Facial Expression Emotions achieved an accuracy of 88.56% and could serve as a new benchmark dataset for future research on fine-grained FER.

Keywords: facial expression recognition (FER); image quality evaluation module; dynamic weight loss function; Swin Transformer; convolutional block attention module (CBAM)



Citation: Jin, Z.; Zhang, X.; Wang, J.; Xu, X.; Xiao, J. Fine-Grained Facial Expression Recognition in Multiple Smiles. *Electronics* **2023**, *12*, 1089. <https://doi.org/10.3390/electronics12051089>

Academic Editor: Jyh-Cheng Chen

Received: 27 January 2023

Revised: 20 February 2023

Accepted: 21 February 2023

Published: 22 February 2023



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

1. Introduction

Facial expressions play a critical role in expressing thoughts and feelings, and thus facial expression recognition (FER) is essential in the field of computer vision. Two research directions in facial expression recognition exist: continuous and categorical [1,2]. In continuous methods, the face image is usually given, and the facial expression is defined in two continuous dimensions; that is, valence and arousal, which can effectively identify subtle differences in expressions with the help of continuous values [3]. Thus, computers can better understand and distinguish differences in facial expressions. However, the drawback of continuous methods is that labeling is more demanding for data annotators and more time-consuming. Therefore, the study of facial expressions based on continuous methods is limited compared to using categorical methods. Their simple operation and short annotation time make categorical methods popular for FER. However, most existing FER datasets [2,4,5] are limited to analyzing only six basic emotion classes (i.e., happy, sad, surprise, anger, fear, and disgust) or seven classes plus an extra neutral emotion, according to Ekman’s theory [6], which is widely used in computer vision but is disadvantageous for fine-grained FER. Most researchers have also conducted research [7–13] on human facial expressions based on Ekman’s theory, while few have focused on exploring the richness and diversity of human emotions.

Human facial expressions take various forms; thus, using only six or seven basic emotion classes is insufficient to cover the current needs of fine-grained FER. Smiling, in

particular, is the most complex expression. In a study of popular FER datasets, the vast majority of FER datasets had a large proportion of “non-happy” expressions in the “happy” class of facial expressions, as shown in Figure 1. This is because there are many types of human smiles, and slight facial changes correspond to different emotions; a special case that is worth studying is deception in expressions during public speeches. To solve this problem, we built a new dataset (named Facial Expression Emotions) that contains six basic smile classes (i.e., guffaw, laugh, beaming smile, qualifier smile, polite smile, and contempt smile) based on Duchenne de Boulogne’s theory [14] of the mechanism of human facial expressions.

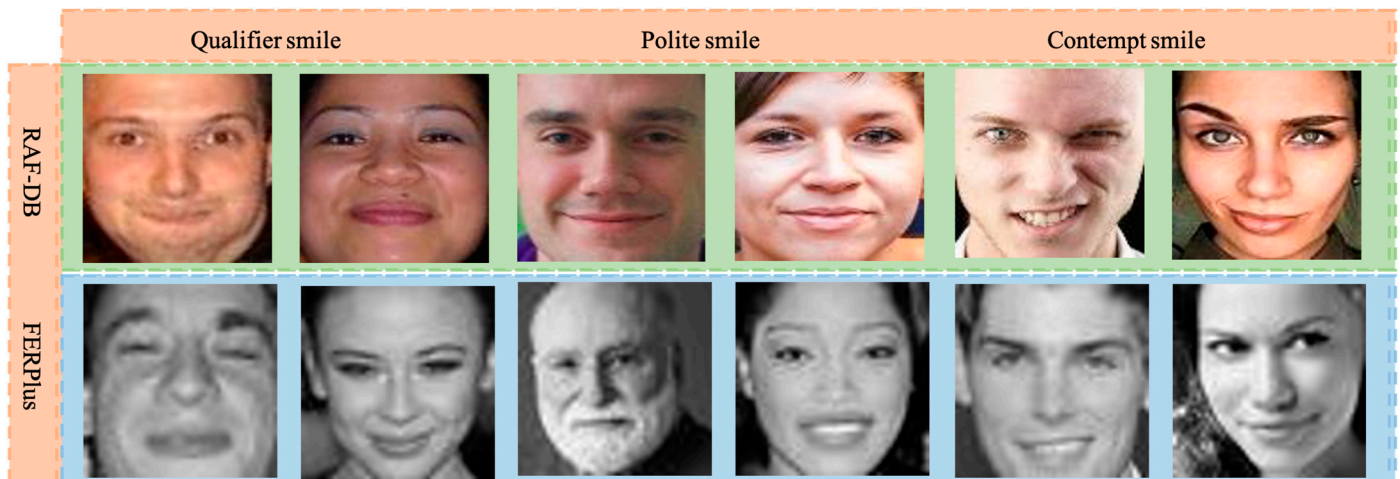


Figure 1. The green images are from the RAF-DB dataset (“happy” class and “non-happy” expressions), and the blue images are from FERPlus dataset (“happy” class and “non-happy” expressions).

In contrast to early FER datasets, the image data of FER datasets [15–18] based on laboratory scene collection typically lack data diversity, exaggerated expressions, and clear positive faces. Currently, motion-blurred and highly degraded images are becoming an indispensable part of datasets, and the effective use of these data has become an important problem. To deal with this, we propose an image quality evaluation module and a dynamic weight loss function to assign different labels to different quality images, focusing on hard yet recognizable samples and discarding unidentifiable images to achieve better classification results in Facial Expression Emotions.

The existing methods based on convolutional neural networks (CNNs) still face challenges [9,10,19–21] in fine-grained FER in-the-wild tasks. When there are different forms of occlusion in facial expressions, such as occlusion of key parts, insufficient profile face information, and different lighting conditions, the CNN is affected [22]. Therefore, feature extraction and recognition methods for expression image features must be further improved [23]. Based on the recent success of the Transformer model in feature extraction and recognition tasks with expression image features, we propose the Smile Transformer network to enhance the local perception of the model and improve the accuracy of fine-grained FER, using the Swin Transformer as a backbone [24]. A convolutional block attention module (CBAM) [25] was designed to focus on important features of the face image and suppress unnecessary regional responses.

Our contributions are summarized as follows:

1. We have created a dataset of human smile expressions, named Facial Expression Emotions, which contains six basic smile classes (i.e., guffaw, laugh, beaming smile, qualifier smile, polite smile, and contempt smile). The dataset can be used for further research on exploring the richness and diversity of human emotions.
2. We have developed an image quality evaluation module that assigns different weights to different complex samples according to their image quality. Then it uses them in

the dynamic weight loss function to dynamically adjust the weights using different image qualities and avoiding emphasizing unidentifiable images, while focusing on hard yet recognizable samples in the loss-function stage.

3. We have proposed the Smile Transformer network, to enhance the local perception of the model and improve the accuracy of fine-grained FER, using the Swin Transformer as a backbone. A CBAM was designed to focus on important features of the face image and suppress unnecessary regional responses. The focus was placed on extracting strong expression correlation features and effectively suppressing background interference.

In summary, our contributions are not only a novel dataset of smiling facial images, which have been gathered from online images and annotated images, but also a method for improving FER in smiling images is proposed. The remainder of this paper is organized as follows: Section 2 presents the recent work on FER datasets and method innovations. Section 3 introduces the Facial Expression Emotions dataset. Section 4 proposes a method that can effectively distinguish different smiles. Section 5 describes the implementation details and the comparative and ablation experiments. Finally, Section 6 summarizes the contributions and shortcomings of our work.

2. Related Work

In the past, face datasets such as CK+ [15] and JAFFE [16] lacked data diversity, exaggerated expressions, and clear positive faces, limiting the development of the industry. Early FERs extracted manual features including grayscale, texture, color, and geometric shape, such as the histogram of oriented gradients (HOG) [26], local binary pattern (LBP) [27], and scale-invariant feature transform (SIFT) [28]. These have been the most used methods, although they were influenced and less generalized due to the interference of external factors (lighting, angle, complex background, etc.), which resulted in serious degradation of the recognition rate [29,30].

In recent years, Li et al. [5] proposed the RAF-DB dataset based on 29,672 real maps of seven classes of facial expressions and 12 classes of composite facial expressions. The dataset contained five accurate landmark locations, 37 landmarks, bounding boxes, race, age range, and gender attribute annotations. Liang et al. [31] proposed extending the original six classes of basic expressions according to Parrott's theory [32], into a more refined FG-Emotions dataset of 33 classes of field facial expressions. This dataset contained a total of 10,371 images and 1491 video clips, used to lay the foundation for further research on fine-grained FER. Wang et al. [33] proposed a large-scale multi-scene FER dataset (FERV39k) with approximately 38,935 video segments subdivided into four main scenarios, which can be subdivided into 22 scenes and seven classic expressions. Chen et al. [34] proposed a method using earlier psycholinguistic research, which selected 135 emotion names from hundreds of English emotion phrases in a prototypicality evaluation analysis. By compiling a sizeable 135-class FER image dataset based on 135 emotion categories, they analyzed the related facial expressions and suggested a follow-up facial emotion recognition framework.

To recognize the significance of facial area occlusion and construct a robust FER, Wang et al. [19] proposed the regional attention network (RAN), which includes a feature extraction module, self-attention module, and relation attention module. The two latter stages have the objective of learning coarse attention weights and refining them in a global context. RAN learns the attention weights for each facial region given a set of facial regions in an end-to-end manner, and then combines their CNN-based features into a small fixed-length representation. This method can improve occlusion performance and pose variation. Farzaneh et al. [20] proposed the use of the deep attentive center loss (DAKL) method to flexibly choose a subset of important features. Using the intermediate spatial feature maps in the CNN as a context, the proposed DAkl included an attention mechanism to estimate the attention weights connected with feature importance. To selectively achieve intraclass compactness and interclass separation for pertinent information in the embed-

ding space, the estimated weights accommodated the sparse formulation of the center loss. Fard et al. [21] proposed an adaptive correlation (Ad-Corre) loss to direct the network toward producing embedded feature vectors, with a high correlation for within-class samples and low correlation for between-class samples. The three parts of Ad-Corre are the feature, mean, and embedding discriminators. Huang et al. [35] proposed a novel facial expression recognition framework with grid-wise attention and a visual transformer (FER-VT) for CNN-based models with two attention mechanisms. To capture the dependencies of various regions from a facial expression image, a grid-wise attention approach was proposed for low-level feature learning, thus regularizing the parameter updating of convolutional filters. A visual transformer attention mechanism in a high-level semantic representation employed a series of visual-semantic tokens to learn the overall representation. Aouayeb et al. [36] achieved competitive results using a squeeze-and-excitation (SE) [37] module in a Vision Transformer.

We summarize the discussed studies in Table 1. In the above innovative datasets, smiling expressions are often incorrectly categorized as “happy”, thus ignoring the complexity of smiling. The datasets in related works were based on the six basic emotions defined by Ekman or their extensions, which cannot effectively distinguish the fine-grained expressions of smiling faces. Although the above innovative methods can improve the accuracy of recognition, they do not consider the problem from the perspective of image quality and loss function. Additionally, to overcome the deficiency of the datasets, we built the Facial Expression Emotions dataset, which contains six basic smile classes. To resolve the shortcomings of previous methods, we proposed the Smile Transformer network, to enhance the local perception of the model.

Table 1. Summary of related works.

Related Work	Methods	Contributions
CK+ [15]	Active appearance models	Laboratory dataset
JAFFE [16]	Gabor coding	Laboratory dataset
RAF-DB [5]	Deep locality-preserving CNN	In-the-wild dataset
FG-Emotions [31]	Multi-scale action unit-based	Fine-grained facial expression recognition in-the-wild dataset
FERV 39k [33]	Four stage candidate clip generation and two-stage annotation workflow	Largescale multi-scene in-the-wild dataset
135-class FER [34]	Pre-trained facial expression embedding and correlation-guided classification	Semantic-rich facial emotional expression recognition in-the-wild dataset
HOG [26], LBP [27], SIFT [28]	Manual design filter	Extract features
RAN [19]	Region attention networks and region generation	Solves real-world pose and occlusion problem
DACL [20]	Sparse center loss and attention network	Improves generalization ability
Ad-Corre [21]	Feature and mean discriminator, embedding discriminator.	Improves generalization ability
FER-VT [35]	Low-level feature learning and high-level feature learning	Solves real-world pose and occlusion problem and improves generalization ability

3. Facial Expression Emotions Dataset

In this section, we introduce Facial Expression Emotions, a new fine-grained facial expression in-the-wild dataset. The most popular categorical method in computer vision is Ekman’s basic emotion theory, with six basic emotions (i.e., happy, sad, surprise, anger, fear, and disgust). However, Ekman’s basic emotion theory has shortcomings for fine-grained FER. Therefore, our approach is based on Duchenne de Boulogne’s theory of the mechanism of human facial expressions, which uses six categories of smiles, accounting for different distinctive features. As shown in Figure 2, we used a hierarchical structure with a root node, secondary nodes (positive, neutral, and negative smiles) [38], and tertiary nodes (i.e., guffaw, laugh, beaming smile, qualifier smile, polite smile, and contempt smile). The process of building the Facial Expression Emotions dataset consisted of three phases: image definition, image collection and preprocessing, and image annotation and inspection.

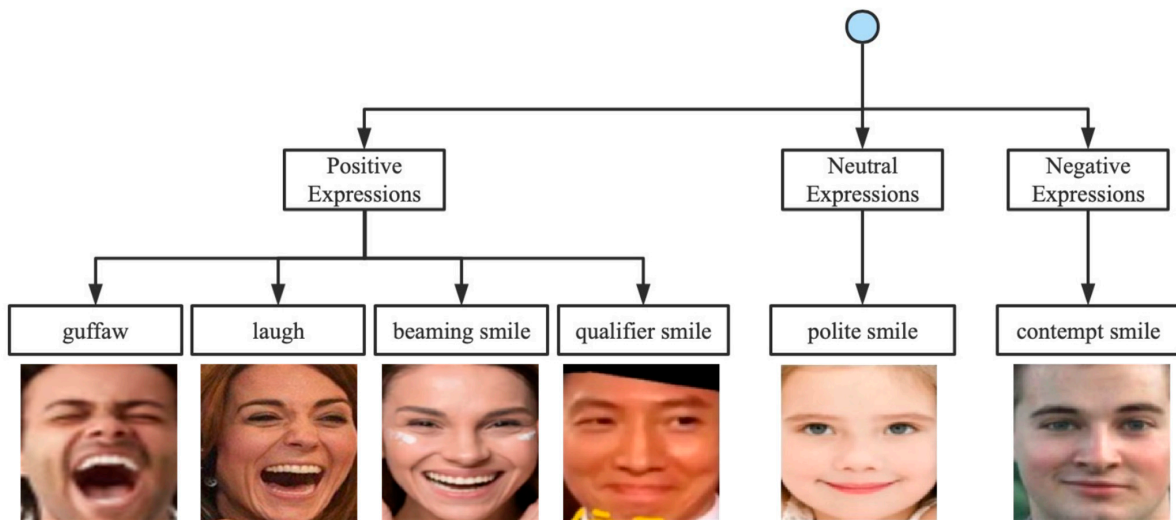


Figure 2. Hierarchical structure adopted in Facial Expression Emotions.

In the image definition phase, 11 experts defined a set of phrases for each category (smile-related keywords combined with gender- and age-related words) and obtained about 50 phrases in English, such as “crazy laughing girl”, “laughing boy”, “middle-aged man smiling contemptuously”, “grandma with a mocking smile”, and “grandpa with a polite smile”. We translated the keywords into nine other languages based on the number of speakers worldwide: Chinese, Spanish, Hindi, Arabic, Malay, French, Bengali, Russian, and Portuguese. Translating English into other languages does not always accurately convey the intended sentiment, as different languages and cultures have their own unique forms of expression. Thus, translation cross-references were obtained from professionals, for accurate search-engine searches.

In the image collection and preprocessing phase, image collection using four search engines (Google, Bing, Yahoo, Baidu) queried approximately 450 emotion-related tags; expressions included only human facial expressions, excluding animation, painting, and other non-human objects. In this phase, because the images were collected using popular search engines, a series of preprocessing processes were carried out considering the different face sizes, positions, and background noise levels, as well as the face detection, face alignment, and image size. The FaceNet [39] face detection framework and the Dlib [40] algorithm were used for detection and face alignment, respectively. The final image was unified as a 224 × 224 image, as shown in Figure 3.

In the image annotation and inspection phase, crowdsourcing services such as Alibaba and Amazon Mechanical Turk provided good options. However, the quality of annotation from crowdsourcing services varied. We hired 11 full-time and part-time annotators from Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences to label the database. A total of 12,400 images were given to these professional data annotators, to label the faces in the images. Images that were categorized in the same smile class by six experts were considered a reliable label; otherwise, the 11 experts re-voted on the category of the image and discarded the image if its category was still difficult to determine. This task was performed by 11 professional data annotators in a month. The associated characteristics of the resulting 11,000 images of Facial Expression Emotions are shown in Table 2.

Table 2. List of associated characteristics for Facial Expression Emotions.

Dataset	Annotation	Training Set	Validation Set	Test Set	Total
Facial Expression Emotions	6 fine-grained class	7856	1572	1572	11,000

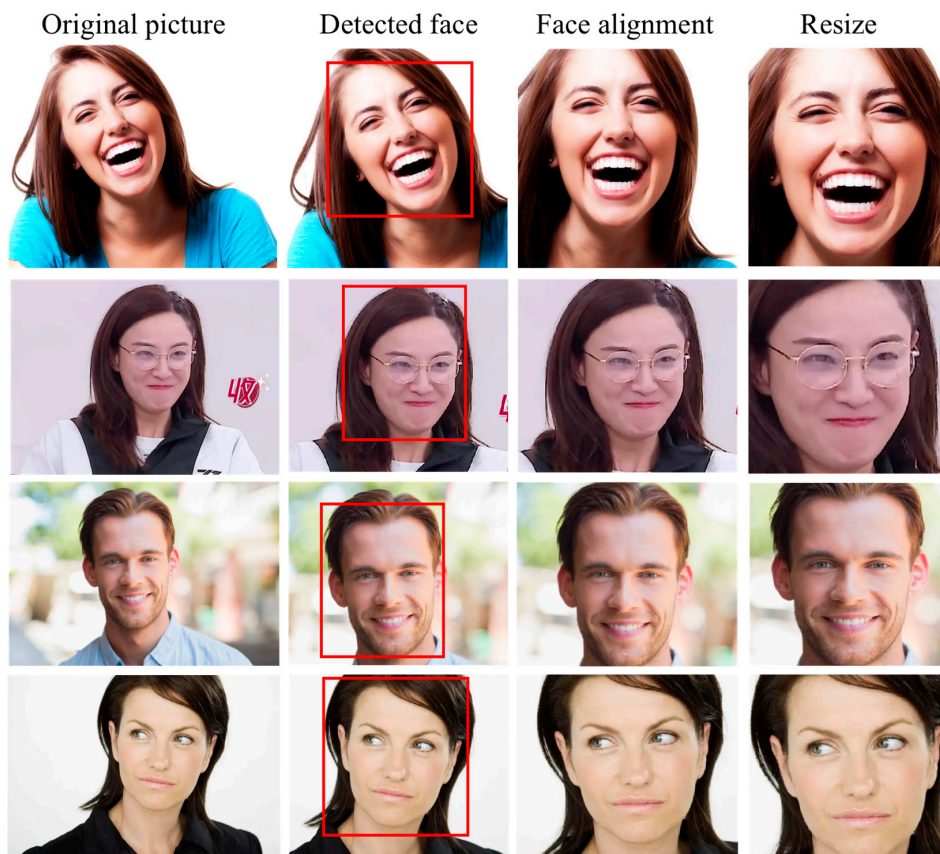


Figure 3. Processing of Facial Expression Emotions.

Figure 4 shows the data volume distribution of the six classes of expressions; it is obvious that our dataset suffers from a class imbalance. To improve this situation, we used flipping, scaling, and cropping to achieve data augmentation.

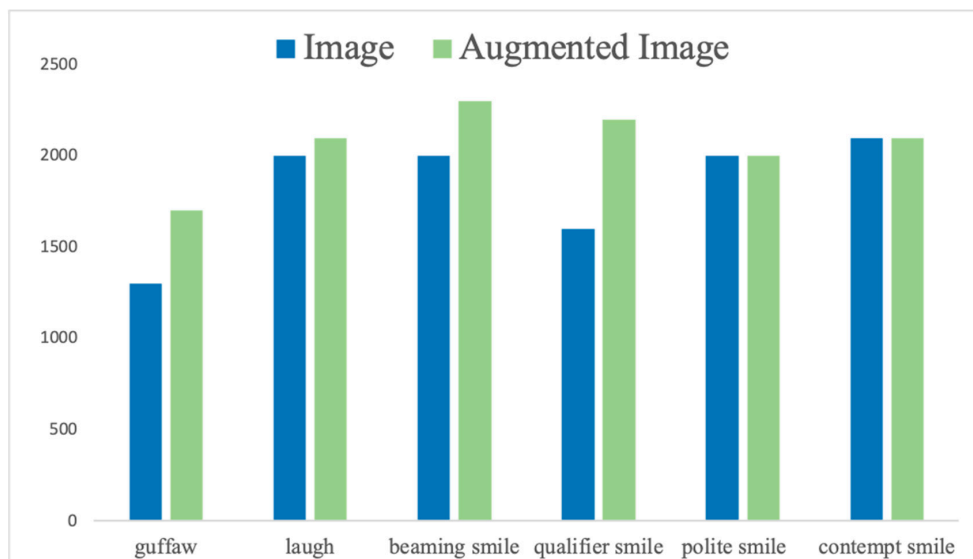


Figure 4. Data distribution over images and augmented image datasets for smile-emotions.

4. Methods

We improved the early FER model in terms of the data preprocessing, backbone network, and loss function. The backbone network is based on the Swin Transformer [24] and

was mainly composed of the patch partition, linear embedding, Swin Transformer, patch merging, and fully connected (FC) layers. The patch partition layer segments images into non-overlapping patches. The linear embedding layer is applied to this raw-valued feature to project it to an arbitrary dimension. With the deepening of the network, the number of tokens is reduced by the patch merging layer, and finally a hierarchical representation is generated. The FC layer plays the role of mapping the learned distributed feature representation onto the sample label space. The remainder of this section introduces the Swin Transformer block, data preprocessing, and loss function. The overall architecture is shown in Figure 5, and we propose a method that can effectively distinguish different smiles.

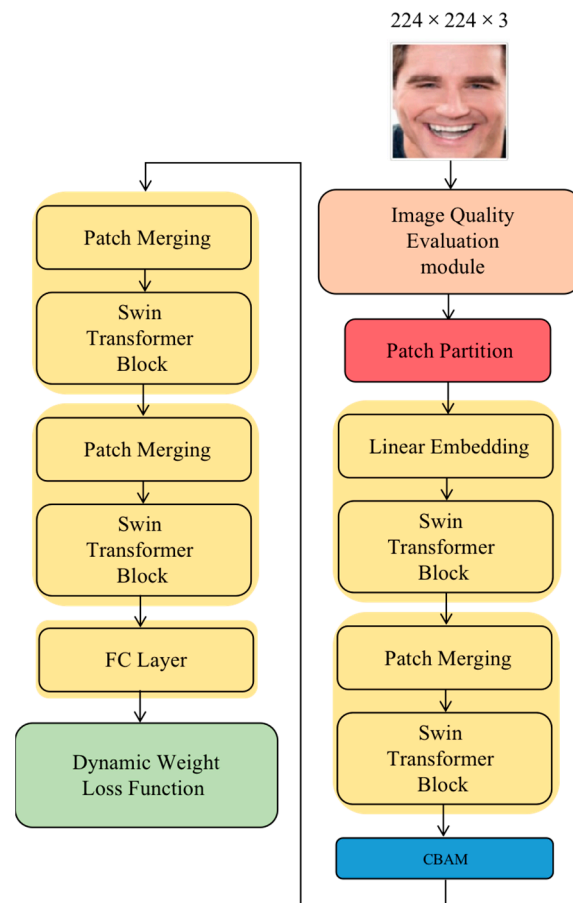


Figure 5. Illustration of the proposed Smile Transformer.

4.1. Data Preprocessing

FER of low-quality images is a challenge, because of the blurring and degradation of images. Image quality is affected by many factors, including noise, brightness, contrast, sharpness, resolution, and hue. Human faces present different images under different facial expressions, poses, and lighting; occasionally, facial expressions are captured in extreme situations, and it is difficult to recognize them. When the facial image quality is low, the FER task becomes infeasible. Figure 6 shows high-quality images, low-quality images, and images with different recognition levels.

Currently, laboratory datasets, such as CK+ [15] and JAFFE [16], have been validated as having over 95% accuracy; nevertheless, FER is primarily intended for use in real-life situations. With the widespread use of cameras, the challenge of FER has shifted to data-rich, low-quality image datasets, and low-quality facial expression images are increasingly becoming an important part of facial expression datasets such as AffectNet and RAF-DB. When low-quality images are excessively degraded, the information related to facial expressions also disappears from the images; thus, the images cannot be recognized. These

unrecognized images are harmful to network training, as the model cannot focus on useful features when there are incorrectly identified features in the images (e.g., hair, occlusions, or image color). Therefore, the model may perform poorly if low-quality images are present in the dataset. To solve the image quality problem, we developed an image quality evaluation module, which mainly consists of overexposed, over-dark, Gaussian-blurred, and object-obscured smiling face images, and generates image quality labels, thus guiding the loss function according to the labels. Based on this, a new loss function, called the dynamic weight loss function, was designed, to assign different weights to different image qualities and to enhance the learning of high-quality complex and low-quality simple samples, as well as to reduce the learning of impossible identification images.





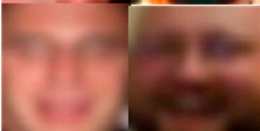
		Easy identification	Hard identification	Impossible identification
High Quality		✓		
			✓	
			✓	
Low Quality			✓	
				✓

Figure 6. Different levels of image recognition with varying image qualities.

4.2. Backbone Network

CNNs are advantageous for extracting deep features and visual aspects but have modelling limitations because of their convolutional structure with low-level semantic information, as shown in Figure 7. A two-layer multilayer perceptron (MLP) [41] with Gaussian error linear unit (GELU) nonlinearity followed a shifted window-based multi-head self-attention (MSA) [42] module in the Swin Transformer block. Each MSA module and MLP underwent layer normalization (LN) in advance; after that, each module also had an residual connection added. In the subsequent Transformer block, window-based multi-head self-attention (W-MSA) and shifted-window multi-head self-attention (SW-MSA) were used. Based on the shifted-window partitioning approach, the successive Swin Transformer blocks can be defined as follows:

$$\hat{t}^l = W - MSA(LN(t^{l-1})) + t^{l-1} \tag{1}$$

$$t^l = MLP(LN(\hat{t}^l)) + \hat{t}^l \tag{2}$$

$$\hat{T}^{l+1} = SW - MSA(LN(t^l)) + t^l \tag{3}$$

$$t^{l+1} = \text{MLP}(\text{LN}(\hat{t}^{l+1})) + \hat{t}^{l+1} \tag{4}$$

where \hat{t}^l represents either W – MSA or SW – MSA; LN represents the LN layer; and t^l represents MLP.

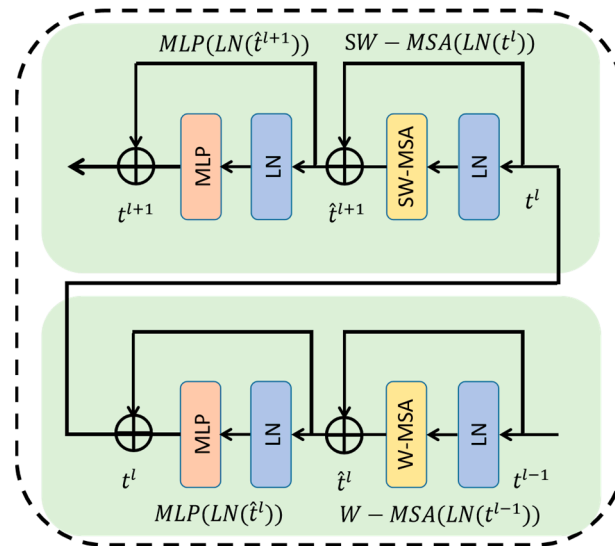


Figure 7. Two successive Swin Transformer blocks.

Moreover, the Swin Transformer is an improved model based on the Transformer, which not only models the native network, focusing on global information, but also uses a shifted-window partitioning approach to realize the connection between different windows, so that the model can better focus on the relevant information of other adjacent windows. The feature interaction of different windows expands the feeling field to a certain extent, thus resulting in a higher efficiency, and greatly reduces the computational complexity compared to the Transformer. We used a CBAM that comprises two parts: a channel attention module (CAM) and a spatial attention module (SAM). The combination of these two modules allowed us to better extract global feature information of faces and strong correlation features, suppress background interference of facial expressions in neural network feature maps, improve the Swin Transformer focus on facial expressions, and partially solve the local occlusion problem.

4.3. Dynamic Weight Loss Function

It is difficult to fully portray the differences in facial expressions with the standard cross-entropy loss training network, and it is also impossible to exploit the information of images with different qualities. Only random guesses can be made for unidentifiable images, which inevitably exist in dataset production and make the results of image annotation unreliable. We proposed a new loss function, to guide model learning by assigning different weights to images of different complexities according to the image quality, expressed as follows:

$$y_1 = -\sum_{i=1}^n \alpha \times p(x_i) \log q(x_i), \tag{5}$$

$$Y_2 = -\sum_{i=1}^n p(x_i) \log q(x_i), \tag{6}$$

$$L = \frac{\text{CEpoch}}{\text{AllEpoch}} \times y_1 + \frac{\text{AllEpoch} - \text{CEpoch}}{\text{AllEpoch}} \times y_2, \tag{7}$$

where AllEpoch is the total number of epochs, CEpoch is the current number of epochs, α is the image quality weight, $q(x_i)$ is the probability of each category, and $p(x_i)$ is the corresponding category value. There was a gradual transition from the standard cross-entropy loss function to a loss function dominated by image quality. This design allowed

the network to make easy identifications in the early stage, enhanced the learning of hard images, and suppressed the learning of impossible to identify images in the later stage.

5. Results

5.1. Dataset

We used our Facial Expression Emotions dataset, which contained more than 11,000 facial images of expressions downloaded from the Internet and professionally annotated, using six basic smile images (i.e., guffaw, laugh, beaming smile, qualifier smile, polite smile, and contempt smile). Facial Expression Emotions contained 7856 images for training, 1572 images for validation, and 1572 images for testing.

5.2. Implementation Details

The framework of the image quality evaluation module was built on a PyTorch GPU using NVIDIA GeForce RTX 2080Ti GPUs. The image quality evaluation module used was ResNet-18 [43], which is a standard CNN. The input and output image sizes were 224×224 pixels. We used a standard stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 0.0005. We augmented the input images on the fly, by extracting random crops.

The implementation environment was the same as above; namely, images with image quality labels were fed into the network. As the facial smile expression recognition domain was close to the FER, we pre-trained the Smile Transformer on RAF-DB, which is a facial expression dataset with 30,000 images. For testing, the central crop of the input image was used. A crop of size 224×224 was obtained from the input images of size 256×256 . For all training cases, data augmentation was used to balance the data and increase data diversity. The input image and patch sizes were set to 224×224 and 3×4 , respectively. An AdamW [44] optimizer was used for 200 epochs, with a batch size of 64, an initial learning rate of 0.0001, and a weight decay of 0.05. At the dynamic weight loss function stage, we focused on hard yet recognizable samples and discarded unidentifiable samples. According to experimental tests, the hyperparameters were set to 0.3, 0.5, and 0.1, for easy, hard, and impossible identification, respectively, to obtain optimal results.

5.3. Recognition Results

To verify the performance of the Smile Transformer, we compared our method with the state-of-the-art open-source methods developed in the past three years on RAF-DB and FERPlus, in terms of accuracy on the Facial Expression Emotions dataset, as shown in Table 3. We plotted the receiver operating characteristic (ROC) curve, as shown in Figure 8

Table 3. Expression recognition performance (in terms of accuracy) of the various methods.

Model	Accuracy (%)
Swin Transformer [24]	80.03%
RAN [18]	84.10%
AD-Corre [21]	84.86%
DAFL [20]	85.34%
FER-VT [35]	86.36%
Smile Transformer	88.56%

On the Facial Expression Emotions wild FER, our Smile Transformer model exhibited a very good performance, with a higher accuracy (88.56%) than FER-VT (86.36%) and DAFL (85.34%). Our ROC curve was better than the other methods. This was achieved because we used the image quality evaluation module to label each image, which enabled the network to learn more detailed features from different quality images. The Smile Transformer was used to enhance the local perception capability of the model, which gave the network a more accurate recognition ability and better recognition results. Concurrently, the standard

Transformer models have more parameters than CNNs. However, the Smile Transformer has a higher recognition accuracy and less parameters than FER-VT Transformer models.

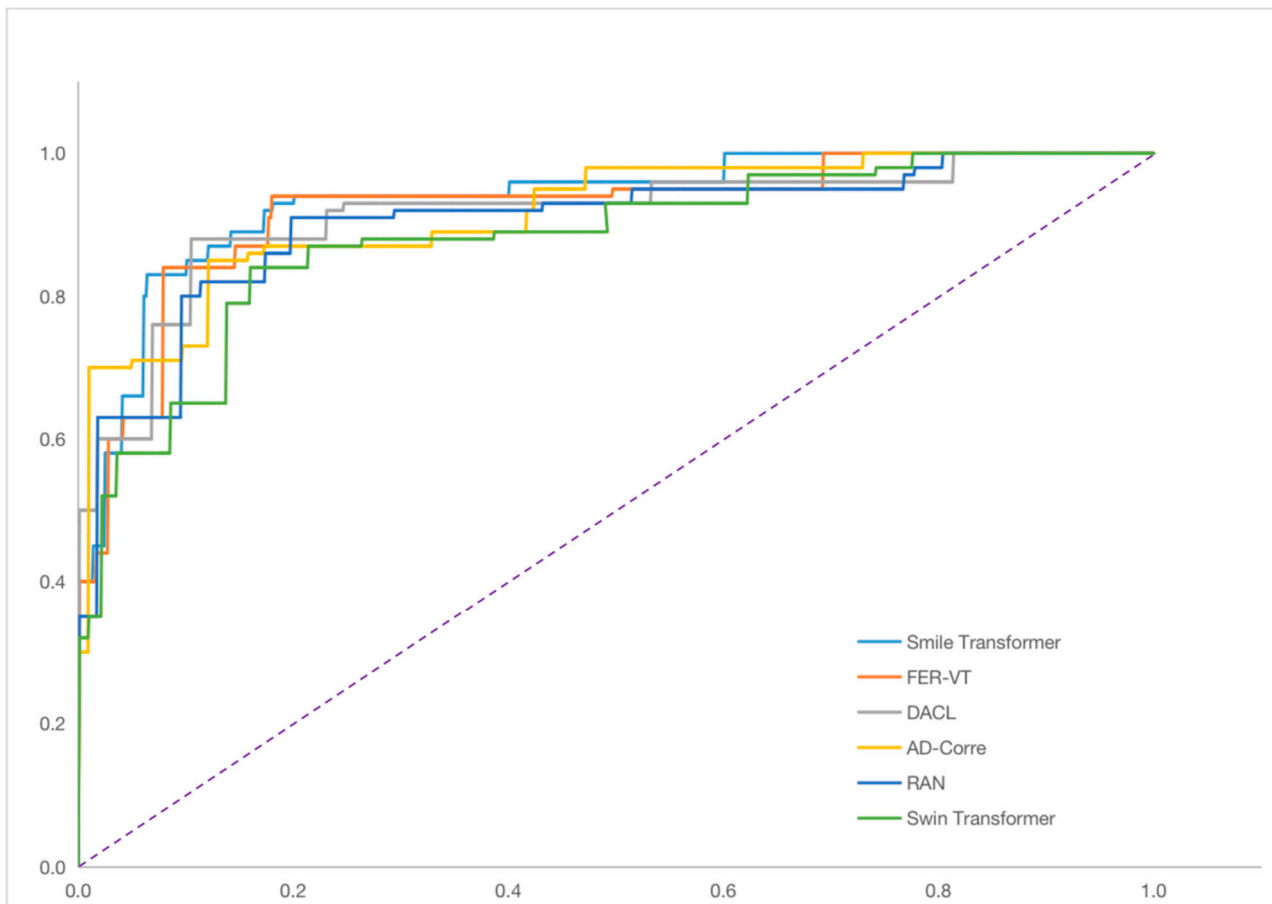


Figure 8. ROC curve of various methods.

To verify the effectiveness of the image quality evaluation module, dynamic weight loss function, and CBAM, we performed ablation experiments on the Facial Expression Emotions dataset using the Swin Transformer as the benchmark model. The experiments were conducted using the same training settings and images with a size of 224×224 , as shown in Table 4. An image quality evaluation module was added to the baseline, and the loss function was modified to improve the recognition accuracy by 6.50%; the recognition accuracy was improved by 2.33% using only the CBAM and by 8.53% when the image quality evaluation module, dynamic weight loss function, and CBAM were used together. The Smile Transformer achieved better performance because the image quality evaluation module assigned different weights to different quality images before training, which better guided the network. Moreover, the model training phase strengthened the attention to global facial features and used a better loss function; thus, it effectively solved difficult images, such as facial occlusions and profile images.

Table 4. Ablation experiment on the Facial Expression Emotions dataset (in terms of accuracy).

Image Quality Evaluation Module and Dynamic Weight Loss Function	CBAM	Facial Expression Emotions
×	×	80.03%
✓	×	86.53%
×	✓	82.36%
✓	✓	88.56%

Owing to the different feature map sizes at different stages of the baseline, the main difference for CBAM being placed at different stages was the different global information and dimensionality represented by the feature maps. To verify the accuracy with the being CBAM placed at different locations, the final accuracy results (when the CBAM was located at different stages) are shown in Table 5. The best accuracy was obtained at Stage 2.

Table 5. Accuracy of CBAM being placed at different stages.

Stage	Accuracy (%)
Stage 1	81.01%
Stage 2	82.36%
Stage 3	81.53%
Stage 4	82.20%

6. Conclusions

To solve the complex problem of smile recognition, we introduced a new benchmark dataset, named Facial Expression Emotions, which consists of six types of fine-grained smile images, to promote research in the field of fine-grained FER. Our Smile Transformer used the Swin Transformer as our baseline for facial expression recognition. The image quality evaluation module assigned different labels to images of different quality; the dynamic weight loss function improved the attention of the network, and CBAM focused on important features of the face image and suppressed unnecessary regions. The experimental results showed that the recognition accuracy on the Facial Expression Emotions dataset reached 88.56%, which was better than that of the other methods. In the future, we will continue our efforts to build a more comprehensive fine-grained smiling face dataset, with more smiling face images and detailed category annotations, to further promote fine-grained FER research.

Author Contributions: Conceptualization, Z.J.; formal analysis and data curation, J.W. and X.X.; supervision, X.Z., J.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ningbo Science and Technology Innovation 2025 Major Project (NO. 2022Z077, No. 2021Z037), by the Key R&D Project of the National Emergency Management Department (NO. 2021XFCX352025).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author, X.Z., upon reasonable request.

Acknowledgments: Thanks to the editors and reviewers for their careful reviewing, and constructive suggestions and reminders.

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

References

- Gunes, H.; Schuller, B. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image Vis. Comput.* **2013**, *31*, 120–136. [[CrossRef](#)]
- Mollahosseini, A.; Hasani, B.; Mahoor, M.H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **2017**, *10*, 18–31. [[CrossRef](#)]
- Kansizoglou, I.; Misirlis, E.; Tsintotas, K.; Gasteratos, A. Continuous Emotion Recognition for Long-Term Behavior Modeling through Recurrent Neural Networks. *Technologies* **2022**, *10*, 59. [[CrossRef](#)]
- Barsoum, E.; Zhang, C.; Ferrer, C.C.; Zhang, Z. Training deep networks for facial expression recognition with crowd-sourced label distribution. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo Japan, 12–16 November 2016; pp. 279–283.
- Li, S.; Deng, W.; Du, J.P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2852–2861.

6. Ekman, P.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* **1971**, *17*, 124. [[CrossRef](#)]
7. Ma, F.; Sun, B.; Li, S. Robust facial expression recognition with convolutional visual transformers. *arXiv* **2021**, arXiv:2103.16854.
8. Li, H.; Sui, M.; Zhao, F.; Zha, Z.; Wu, F. MVT: Mask vision transformer for facial expression recognition in the wild. *arXiv* **2021**, arXiv:2106.04520.
9. Wen, Z.; Lin, W.; Wang, T.; Xu, G. Distract your attention: Multi-head cross attention network for facial expression recognition. *arXiv* **2021**, arXiv:2109.07270.
10. Savchenko, A.V. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In Proceedings of the 2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, Serbia, 16–18 September 2021; pp. 119–124.
11. Vo, T.H.; Lee, G.S.; Yang, H.J.; Kim, S.H. Pyramid with super resolution for in-the-wild facial expression recognition. *IEEE Access* **2020**, *8*, 131988–132001. [[CrossRef](#)]
12. Zhang, Y.; Wang, C.; Ling, X.; Deng, W. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022*; Springer: Cham, Switzerland, 2022; pp. 418–434.
13. Zhang, Y.; Wang, C.; Deng, W. Relative Uncertainty Learning for Facial Expression Recognition. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 17616–17627.
14. Duchenne, G.B.; de Boulogne, G.B.D. *The Mechanism of Human Facial Expression*; Cambridge University Press: Cambridge, UK, 1990.
15. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z. 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, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
16. Lyons, M.; Akamatsu, S.; Kamachi, M.; Gyoba, J. Coding facial expressions with gabor wavelets. In Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 200–205.
17. Wang, W.; Sun, Q.; Chen, T.; Cao, C.; Zheng, Z.; Xu, G.; Qiu, H.; Fu, Y. A fine-grained facial expression database for end-to-end multi-pose facial expression recognition. *arXiv* **2019**, arXiv:1907.10838.
18. Valstar, M.; Pantic, M. Induced disgust, happiness and surprise: An addition to the mmi facial expression database. In Proceedings of the 3rd International Workshop on EMOTION (Satellite of LREC): Corpora for Research on Emotion and Affect, Valletta, Malta, 23 May 2010; p. 65.
19. Wang, K.; Peng, X.; Yang, J.; Meng, D.; Qiao, Y. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Trans. Image Process.* **2020**, *29*, 4057–4069. [[CrossRef](#)]
20. Farzaneh, A.H.; Qi, X. Facial expression recognition in the wild via deep attentive center loss. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 2402–2411.
21. Fard, A.P.; Mahoor, M.H. Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. *IEEE Access* **2022**, *10*, 26756–26768. [[CrossRef](#)]
22. Chen, Z.; Huang, D.; Wang, Y.; Chen, L. Fast and light manifold CNN based 3D facial expression recognition across pose variations. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 229–238.
23. Kansizoglou, I.; Bampis, L.; Gasteratos, A. Deep feature space: A geometrical perspective. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 6823–6838. [[CrossRef](#)]
24. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
25. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
26. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
27. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [[CrossRef](#)]
28. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [[CrossRef](#)]
29. Kansizoglou, I.; Bampis, L.; Gasteratos, A. An active learning paradigm for online audio-visual emotion recognition. *IEEE Trans. Affect. Comput.* **2019**, *13*, 756–768. [[CrossRef](#)]
30. Cai, J.; Meng, Z.; Khan, A.S.; Li, Z.; O'Reilly, J.; Han, S.; Liu, P.; Chen, M.; Tong, Y. Feature-level and model-level audiovisual fusion for emotion recognition in the wild. In Proceedings of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, 28–30 March 2019; pp. 443–448.
31. Liang, L.; Lang, C.; Li, Y.; Feng, S.; Zhao, J. Fine-grained facial expression recognition in the wild. *IEEE Trans. Inf. Forensics Secur.* **2020**, *16*, 482–494. [[CrossRef](#)]
32. Parrott, W.G. *Emotions in Social Psychology: Essential Readings*; Psychology Press: Philadelphia, PA, USA, 2001.

33. Wang, Y.; Sun, Y.; Huang, Y.; Liu, Z.; Gao, S.; Zhang, W.; Ge, W.; Zhang, W. FERV39k: A Large-Scale Multi-Scene Dataset for Facial Expression Recognition in Videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20922–20931.
34. Chen, K.; Yang, X.; Fan, C.; Zhang, W.; Ding, Y. Semantic-Rich Facial Emotional Expression Recognition. *IEEE Trans. Affect. Comput.* **2022**, *13*, 1906–1916. [[CrossRef](#)]
35. Huang, Q.; Huang, C.; Wang, X.; Jiang, F. Facial expression recognition with grid-wise attention and visual transformer. *Inf. Sci.* **2021**, *580*, 35–54. [[CrossRef](#)]
36. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
37. 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.
38. Trougakos, J.P.; Jackson, C.L.; Beal, D.J. Service without a smile: Comparing the consequences of neutral and positive display rules. *J. Appl. Psychol.* **2011**, *96*, 350. [[CrossRef](#)] [[PubMed](#)]
39. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNET: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
40. Kazemi, V.; Sullivan, J. One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1867–1874.
41. Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **1989**, *2*, 359–366. [[CrossRef](#)]
42. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, 5998–6008.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
44. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.

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