A channel attention and feature manipulation network for facial expression recognition

Zixin Guo[1]**, and Ruizhi Yang**[2]

[1]Department of Computer Science, University of Toronto, Toronto M5S 2E4, Canada
[2]International School, Beijing University of Posts and Telecommunications, Haidian, Beijing 100876, China

zixin.guo@mail.utoronto.ca

Abstract. Facial expression conveys a variety of emotional and intentional message from human beings, and automated facial expression recognition (FER) has become an ongoing and promising research topic in the field of computer vision. However, the primary challenge of FER is learning to discriminate similar features among different emotion categories. In this paper, a hybrid architecture using Efficient Channel Attention (ECA) residual network ResNet-18, and feature manipulation network is proposed to tackle the above challenge. First, the ECA residual network effectively extract input features with local cross-channel interaction. Then, the feature decomposition network (FDN), feature reconstruction network (FRN) modules are added to decompose and aggregate latent features for enhancing the compactness of intra-category features and discrimination of inter-category features. Finally, an expression prediction network is connected to FRN to draw the final expression classification result. To examine the efficacy of the suggested approach, the model is trained independently using in-the-lab (CK+) and in-the-wild (RAF-DB) datasets. Several important evaluation metrics such as confusion matrix, Grad-CAM are reported, and the ablation study is conducted for demonstrating the efficacy and interpretability of the proposed network. It achieves the state-of-the-art accuracy compared to the existing facial recognition work, at 99.70% in CK+ and 89.17% in RAF-DB.

Keywords: facial expression recognition, ResNet-18, efficient channel attention network, feature decomposition network, feature reconstruction network, CK+, RAF-DB.

1. Introduction

Facial expression is one of the simplest ways to convey feeling and sense between human beings. Having a reliable automated facial expression recognition system would be beneficial for advancing many other active research fields, such as human computer interaction (HCI).

With the advancement in computer vision and deep learning, adapting convolutional neural network (CNN) and its variant to the field of FER has shown to be effective on extracting features of input image and classifying facial expression into categories. Still, there exists a huge challenge for improving the model's accuracy – high similarities across different categories, e.g., as shown in the Figure 1, both happy and neutral expressions have closing eyes; both happy and surprise expressions have opening mouth; both sad and disgusting expressions have frowning brows. Learning with such similarities often results in undesirable results as it disturbs the discriminability of the model.

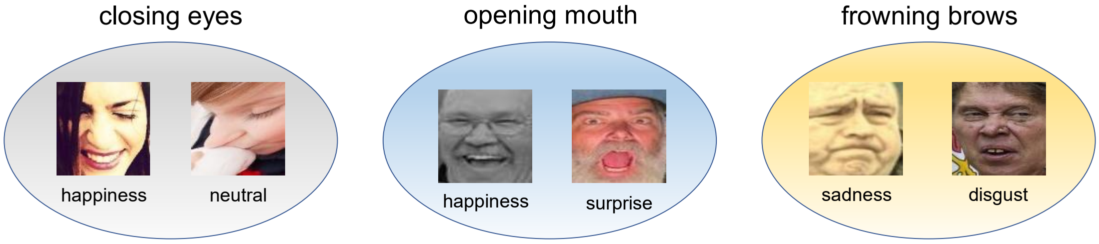

Figure 1. Examples for high similarity between different expressions.

In this paper, an end-to-end hybrid network composed of an efficient channel attention (ECA) [1] residual network ResNet-18 [2], and a Feature Decomposition and Reconstruction Learning (FDRL) [3] network is proposed to tackle the situation. Specifically, our proposed is combined of two sections. The first section is a feature extractor in which the combined ECA and ResNet-18 serve as the backbone to extract the related features. The second section is the FDRL Network, which is consisted of Feature Decomposition Network (FDN), Feature Reconstruction Network (FRN) and Expression Prediction Network (EPN) [3]. FDRL section is used to extract the latent features and learn the characteristics of internal relevance perception and the correlation between the latent features. For the optimization of the model, the joint loss function is composed of several losses, including classification loss, compactness loss, distribution loss, and weight penalty.

As a result, our hybrid model achieves state-of-the-art performance in CK+ dataset [4] and RAF-DB [5] dataset, at 99.70% and 89.17%, respectively. To further reflect the performance and interpretability of the model, confusion matrix is constructed to evaluate the expression prediction accuracy on each category and Grad-CAM [6] is embedded to check salient regions on the activation map for prediction as well. Moreover, the ablation study is conducted by holding one sub-module constant at a time. The experiment is conducted by removing ECA and FDRL, training them separately on both two datasets. and evaluating the results. The results from the ablation study show that both ECA and FDRL jointly contribute to our hybrid model in classifying facial expression as our proposed hybrid model achieves the best accuracy.

2. Related work

Deep learning methods have been widely used for facial expression recognition. The first Convolutional Neural Network (CNN), LeNet [7] was designed to solve MNIST handwritten numeral recognition. The biggest contribution for LeNet was defining the fundamental organization of the convolutional, pooling, and fully connected (FC) layers. Nevertheless in fact, due to the low computing power at the time, CNN, which needs a large amount of computing, can be substituted by other algorithms that requires less computational complexity, such as Support Vector Machine (SVM) [8]. AlexNet [9] is a CNN network with 8 layers of depth, including 5 convolutional layers and 3 fully connected layers. It uses dropout to decrease the overfitting problem [10], ReLU to introduce non-linearity [11], and data augmentation techniques, to make deep learning methods in CNN popular again. Two years later, VGGNet [12] was proposed, which proved that the increasing depth of the network can affect the performance of the result. In the VGG network, each hidden layer utilizes the output of the previous convolutional layers, which are then used to extract more complex features. It turns out that it became the state-of-the-art model compared to previous algorithm at that time. In 2015, a network with landmark innovation, Residual Neural Network from ResNet, was proposed in which many residual blocks are stacked to form a deep network. The residual block not only greatly decreases the problem of gradient vanishing, but also gives a promising improvement to the accuracy of the model. In recent years, a large portion of CNN approach for FER still use the ResNet as the backbone, as it performs exceptionally well on extracting fine-grained features.

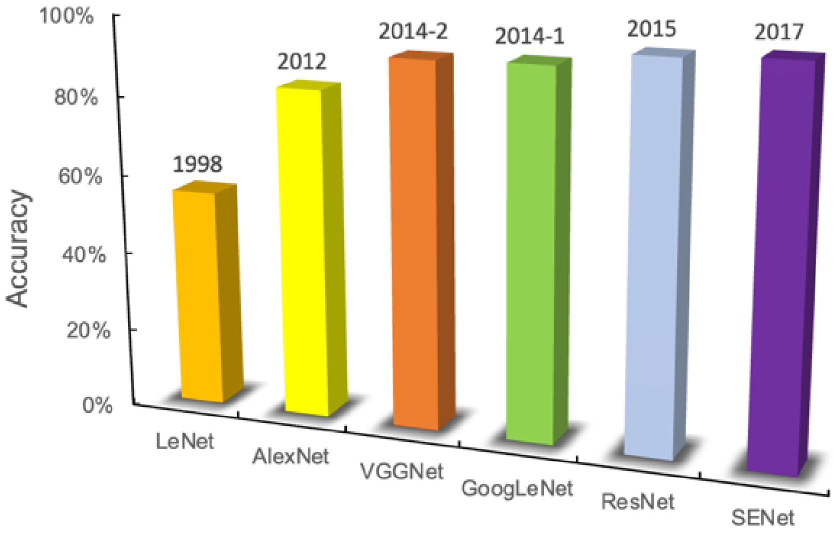

Figure 2. The development of CNN from LeNet to SENet.

CNN used in facial expression detection is driving researchers to go further into a hybrid network to boost model performance. By incorporating information channel attention into convolution blocks in recent years, SENet [13] has generated a lot of interest and shows significant promise for performance improvement in feature extraction. However, with the higher precision and complexity of the model, the computation complexity rises as well. As a result, Efficiency Channel Attention (ECA) was proposed, which focuses on the combination of channel attention and spatial attention. It turns out that ECA achieves a competitive performance while preserving low model complexity.

3. Method

This section introduces our proposed network, an end-to-end framework that improves our baseline model Feature Destruction and Reconstruction Learning (FDRL) by incorporating extra channel-wise attention mechanism (ECA) and additional loss function. An overview framework of our method is illustrated in figure 3.

3.1. Model overview

As shown in figure 3, our model can be roughly divided into two sections, Feature Extractor (Sec. 3.2) and FDRL (Sec. 3.3). Given a batch of normalized images [14], the Feature Extractor learns a mapping from the input to extracted features. Specifically, a residual network ResNet-18 with Efficient Channel Attention (ECA) is used as our backbone to perform the interaction among different channels and efficiently extract the relationship among them. FDRL constitutes the remaining of the model. It is consisted of three portions, feature decomposition module (FDN), feature reconstruction module (FRN) and facial expression prediction (EPN). The features extracted by the previous ECANet (ECA+ResNet-18) will be fed into FDRL. In FDRL, the extracted image features are destructed by FDN, and reconstructed by FRN. Consequently, the output of FRN is fed into the EPN module, which produces a series of probabilities of expression classification, i.e., seven emotion categories.

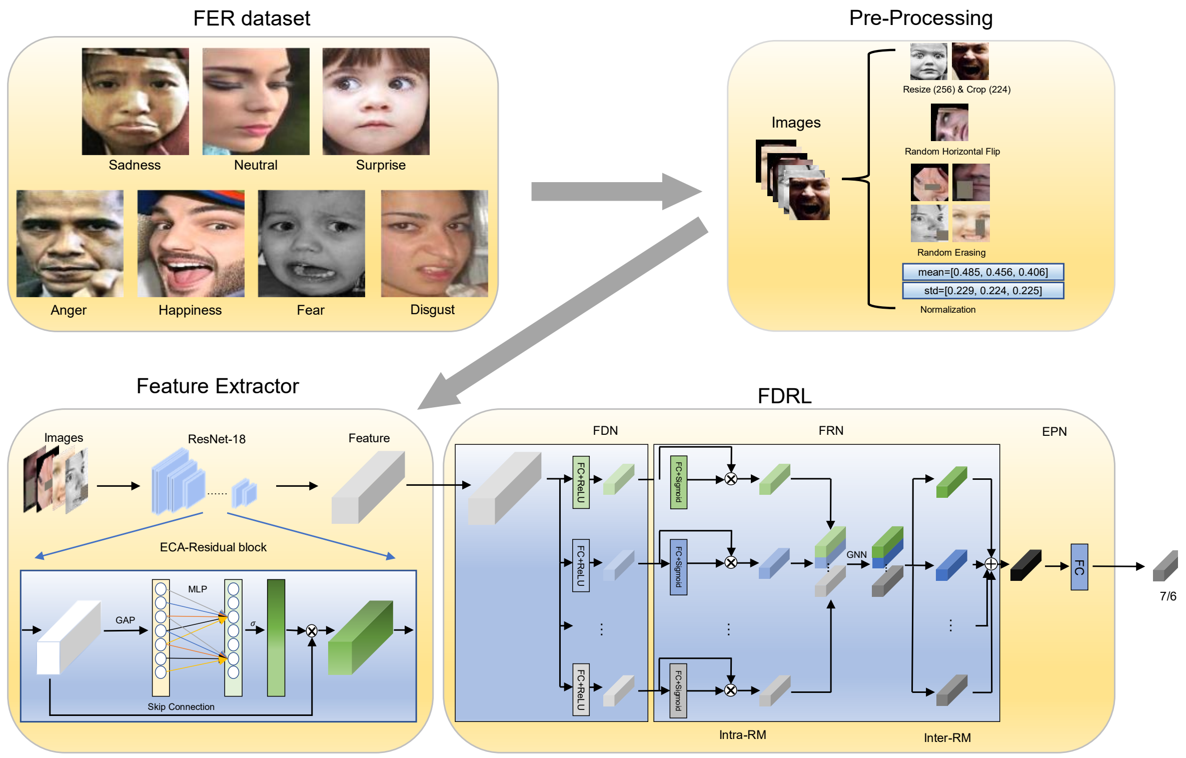

Figure 3. The structure of our model. Feature extractor is the backbone, mainly used for feature extracting; FDRL [3] is consisted of FDL, FRL and EPN, mainly used for features processing and probability prediction.

3.2. Feature extractor

The vanilla Convolutional Neural Network (CNN), the residual network ResNet-18, and an integrated Efficiency Channel Attention (ECA) module make up this component.

The main purpose of the Feature Extractor is to extract the facial features of the input images. First, ResNet-18 has the skip connection in "residual block", which will help to decrease the vanishing gradient problem caused by the huge depth of the network, so it can utilize the good performance of very deep CNN in extracting features. Second, The Efficient Channel Attention (ECA) is also added into our model to introduce attention. In this way, the model explores the interaction of channels and assigns different importance (weight) to different channels. The detailed design and analysis are elaborated in the following subsections.

3.2.1. Efficient channel attention aggregation. Figure 4 illustrates how the Efficient Channel Attention (ECA) module is implemented into our suggested model, which has fewer parameters but significantly improves performance. By breaking down the channel attention module in SENet, avoiding dimensionality reduction is critical for learning channel attention and that effective cross-channel interactions may sustain performance while drastically lowering model complexity. Consequently, 1D convolution may be effectively used to perform the local cross-channel interaction method that does not require dimensionality reduction. A technique for adaptively selecting the size of 1D convolution kernels is also included in ECA in order to assess the extent of local cross-channel interactions.

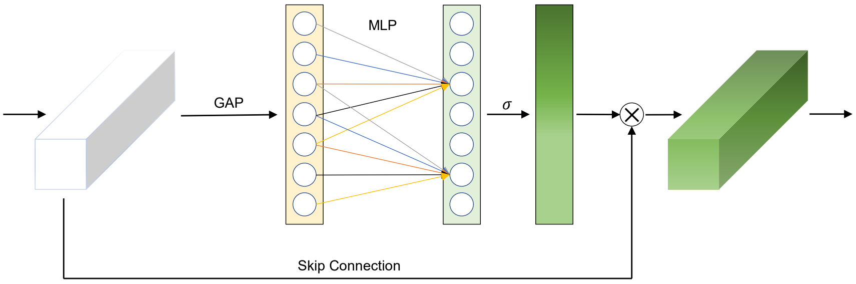

Figure 4. The structure of efficient channel attention (ECA) [1].

3.2.2. Residual backbone. ResNet-18 is a residual-based network that is formed by stacking a lot of tiny residual layer. Each residual layer adds skip connections (identity mapping) to the convoluted features, which is beneficial for optimization by resolving the gradient explosion or gradient vanishing in process of back propagation [15]. As a result, it is shown to alleviate the degradation problem of representation power brought by deepening the network.

3.3. Features decomposition & reconstruction learning (FDRL)

The FDRL is composed of three submodules, including feature decomposition, reconstruction, and facial expression prediction. The feature decomposition module aims to disentangle input features into M latent features in a way that each latent feature of all input samples is clustered in their corresponding centroid. In the reverse process, the feature reconstruction consists of both intra-relationship and inter-relationship modeling of the decomposed latent features, which ensures the extracted features are discriminative. The former one models the relation of same latent feature across all input samples, which learns a weight for each latent feature that determines its contribution of the expression prediction result, whereas the latter one learns the relation of different latent features in the same input sample via a Graph Neural Network (GNN) [16, 17]. Subsequently, the computed features in both intra-relation and inter-relation networks are aggregated and feed into the facial expression prediction module, which only consists of multilayer perceptron for predicting the probability of expression classification.

3.3.1. Feature decomposition. The feature decomposition module maps the input, denoted as $x_i \in \mathrm{R}^{p=512}$, obtained from ECA module to latent features in M subspaces through fully connected layers. Specifically, the "j-th" latent feature of "i-th" input sample is computed as:

$$\mathbf{I}_{i,j} = \sigma_1\left(\mathbf{W}_{d_j}^T \mathrm{x}_i\right) \in \mathbb{R}^{D=128} \tag{1}$$

Then, for the same latent feature of each input sample, a centroid under the feature is clustered iteratively and it is denoted as c_j. To make the latent feature distributions represent different regions of the face, the compactness loss between latent features and corresponding centroids is introduced.

3.3.2. Feature reconstruction. During the process of feature reconstruction, the relationship of both intra-features and inter-features are learned to explore the subtle difference of the latent features between various expressions.

For the modelling of intra-features, the weight of each latent feature of **i**-th input is computed by passing it to a FC layer and a sigmoid function to have a value between 0 and 1 as $\boldsymbol{\alpha}_{i,j} = \sigma_2(\mathbf{W}_{S_j}^T \mathbf{l}_{i,j})$. T.en, the L1 norm of each $\boldsymbol{\alpha}_{i,j}$ is calculated to serve as a weight importance of **j**-th feature of **i**-th input sample. To better remedy the disturbance caused by similar local patterns of different expressions, a distribution loss is proposed to ensure the same features of different samples to be as close as possible. Thus, the final intra-feature for the **i**-th input is computed as: $\mathbf{f}_{\mathrm{i,j}} = \alpha_{\mathrm{i,j}}\mathbf{l}_{\mathrm{i,j}}$.

For the modelling of inter-features, an undirected graph neural network is constructed to learn $\mathbf{f}_{i,j}$ through message passing and aggregating. Specifically, the vertex information is computed by firstly applying FC to the previously obtained intra-feature and a ReLU as the activation function.

$$\mathbf{g}_{i,j} = \sigma_1\left(\boldsymbol{W}_{e_j}^T \mathbf{f}_{i,j}\right) \tag{2}$$

Subsequently, the relation importance $\omega(j,m)$ of different vertices $\mathbf{g}$ are modelled by passing them to a similarity matrix S and a tanh as the activation function.

$$\omega_i(j,m) = \begin{cases} \sigma_3\left(S(\mathbf{g}_{i,j}, \mathbf{g}_{i,m})\right) & j \neq m \\ 0 & j = m \end{cases} \tag{3}$$

Hence, the inter-feature for $\mathbf{i}$-th input sample is obtained by:

$$\hat{\mathbf{f}}_{i,j} = \sum_{m=1}^{M} \omega_i(j,m)\mathbf{g}_{i,m} \quad \text{for } j = 1,2,\dots, M. \tag{4}$$

The resulting reconstructed $\mathbf{j}$-th feature for $\mathbf{i}$-th input is computed as a convex linear combination of $\mathbf{j}$-th intra-feature and $\mathbf{j}$-th inter-feature:

$$\mathbf{y}_{i,j} = \delta\mathbf{f}_{i,j} + (1-\delta)\hat{\mathbf{f}}_{i,j} \quad \text{for } j = 1,2,\dots, M. \tag{5}$$

Finally, M reconstructed features are aggregated to obtain the final representation for $\mathbf{i}$-th input sample

$$\mathbf{y}_i = \sum_{j=1}^{M} \mathbf{y}_{i,j} \tag{6}$$

where $\mathbf{y}_\mathrm{i} \in \mathbb{R}^\mathrm{D}$represents the expression feature for i-th face image.

3.3.3. Expression prediction. Expression prediction module is used to perform the final classification of the facial expression. It is implemented by a simple Multilayer Perceptron (MLP) [3] that maps from features of dimension D to 7, where each entry of the output represents a probability being one of the seven emotions.

3.3.4. Joint loss function. We used the following defined joint loss function in this paper, $L = L_{cls} + \lambda_1 L_C + \lambda_2 L_D + \lambda_3 L_W$ where $\lambda_1, \lambda_2, \lambda_3$ are the coefficients of the loss terms. L_{Cls} is the classification loss, which is opted as the cross-entropy loss between ground truth and predicted expression category. L_C is the compactness loss for feature destruction, which is defined as:

$$\mathrm{L_C} = \frac{1}{\mathrm{N}}\sum_{i=1}^{N}\sum_{j=1}^{M}\left\| l_{i,j} - mathbfc_j \right\|_2^2 \tag{7}$$

where N is the batch size, M is the number of latent features. L_D is the distribution loss for intra-feature modelling:

$$\mathrm{L_D} = \frac{1}{\mathrm{N}}\sum_{i=1}^{N}\left\| \boldsymbol{w}_i - \boldsymbol{w}_{k_i} \right\|_2^2 \tag{8}$$

where N is the batch size, $\boldsymbol{w}_i = [\alpha_{i,1}, \alpha_{i,2}, \dots, \alpha_{i,M}]$ and $\boldsymbol{w}_{k_i} \in \mathrm{R}^M$ represents the class centroid corresponding to the k_i-th expression category. $\mathrm{L_W}$ is the L2 penalty for all parameter weights:

$$\mathrm{J} = \mathrm{J_0} + \frac{\lambda}{2m}\|w\|_2^2 \tag{9}$$

4. Experiment setting

4.1. Dataset description & pre-processing

In this study, we employ the Extended Cohn-Kanade (CK+), RAF-DB, and MS-CelebA [18] databases. 327 video clips from regulated lab conditions are included in CK+. The training set and test set were built using the three peak expression frames from each expression sequence, yielding a total of 981 pictures. RAF-DB is a real-world Facial Expression Recognition (FER) database that has 30,000 photos that 40 trained human labellers have annotated with simple or complex emotions. In our experiment, photographs containing six fundamental expressions and one neutral expression are employed. 3,068 photos are used for testing and 12,271 images are used for training in RAF-DB. The CUHK's MS-

CelebA (also known as CelebA) open data set includes 202,599 photos of 10,177 celebrity identities, each with a size of 178×218 and with clear labels. This dataset exhibits various facial attribution, such as eyes, nose, and mouth, which serves as a prior guide for training our model's feature extractor. Hence, the ECANet backbone network (ECA+ResNet-18) of our proposed network is pretrained on CelebA to reduce the training time and improve the accuracy of the trained model using optimized weights.

During training process, both images from CK+ and RAF-DB are randomly cropped to the size 224×224, and then being transformed with random horizontal flip. Following it, a random rotation and a random crop are applied to the previously transformed data. Then, the normalization is applied to better decrease the computational complexity. At the last step, random erasing is applied to the dataset, which is shown to be effective for avoiding data overfit.

During validation process, the pre-processing is relatively easier than that of training process. For the original data which has been initially crop to 256×256, we perform a central crop to the size 224×224, and then apply the normalization.

As shown in figure 5, the augmented training data are listed as seven classes.

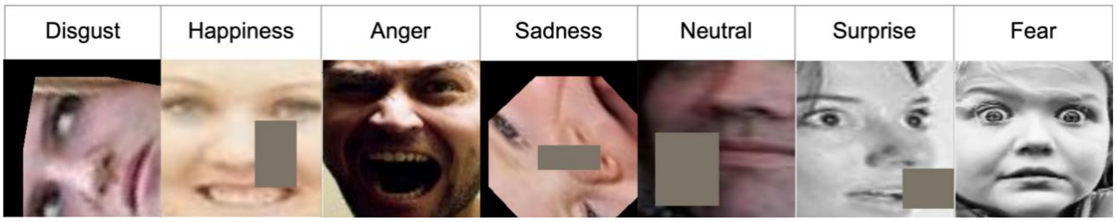

Figure 5. The augmented training data.

4.2. Implementation detail

In both training and testing stages of data augmentation, the images are all converted to the size 224×224. Images in CK+ and RAF-DB dataset are fed into the model by the method Batch Normalization (BN) and the batch size is 32. We train our baseline model in NVIDIA GeForce RTX 3080 for 60 epochs in both RAF-DB and CK+ dataset. The optimizer we choose is Adam and the initially learning rate is 0.0001 and a 0.1-factor decline every 7 epochs. Apart from these, the baseline model (ResNet-18+FDRL) has the same set of hyperparameter settings. In addition, we utilize 10-fold cross-validation for both datasets to evaluate the model more effectively and avoid experimental contingency.

5. Results and discussion

In this section, we provide more detailed about the proposed model performance in our experiment, including overall performance in sec. 5.1 and ablation study in sec. 5.2.

5.1. Overall performance

5.1.1. Model performance and comparison with state-of-the-art. Our proposed models follow the 10-fold cross-validation protocol, as shown in Figure 6, for RAF-DB dataset, the training accuracy curve is smooth and eventually reaches convergence easily. However, there is a crossover between training curve and testing curve on about 26 epochs, which means there exists slight overfit problem; for CK+ dataset, even though the performance on validation set fluctuates in a huge degree, as it may be due to less available samples or a large learning rate, both the validation accuracy and loss converge and match to the training performance.

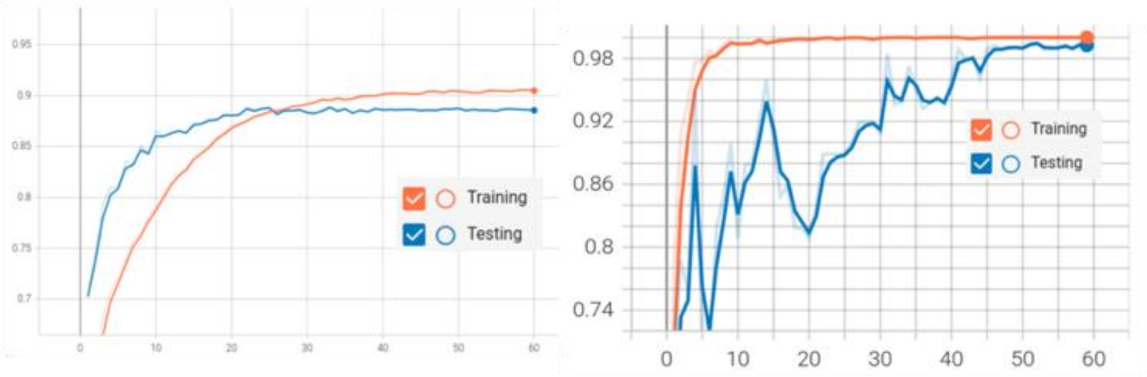

Figure 6. The accuracy curve of training and testing process (left: RAF-DB; right: CK+).

As shown in Table 1, the highest accuracy achieved by our model is 89.13% on RAF-DB dataset and 99.71% on CK+ dataset. Therefore, the experiment shows that our hybrid model competes with other state-of-the-art models.

Table 1. 7 expression categories comparison study with other cutting-edge work on both CK+ (left) and RAF-DB (right) dataset, bolded results are the most effective. (7) and (6) indicate that CK+ uses seven expression categories and six expression categories, respectively.

CK+		RAF-DB	
Method	Accuracy (%)	Method	Accuracy (%)
PPDN [19]	97.30 (6)	IACNN [20]	-
IACNN [20]	95.37 (7)	DLP-CNN [21]	84.13
DLP-CNN [21]	95.78 (6)	IPA2LT [22]	86.77
IPA2LT [22]	92.45 (7)	SPDNet [23]	87.00
DeRL [24]	97.37 (7)	RAN [25]	86.90
FN2EN [26]	98.60 (6)	SCN [27]	87.01
DDL [28]	99.16 (7)	DDL [28]	87.71
Baseline	99.54 (7)	Baseline	88.47
FDRL-ECANet (proposed)	**99.70 (7)**	FDRL-ECANet (proposed)	**89.13**

5.1.2. Visualization. As shown in figure 7, the confusion matrix directly reflects the relationship between true label and predicted label. For CK+ dataset, the diagonal line of its confusion matrix is closed to 1 and the other area is closed to zero, meaning that the overall classification accuracy reaches to nearly 100%. For RAF-DB dataset, the diagonal line of its confusion matrix is also the darkest of the whole area. But compared to the confusion matrix of CK+ dataset, the True Prediction (TP) block is shallower, which has lower accuracy than that. It should be noted that our model still fails at some hard negative examples. For example, the “Fear” class only achieves an accuracy around 65%. But overall, our model achieves a high accuracy in both CK+ (99.706%) and RAF-DB (89.13%) dataset.

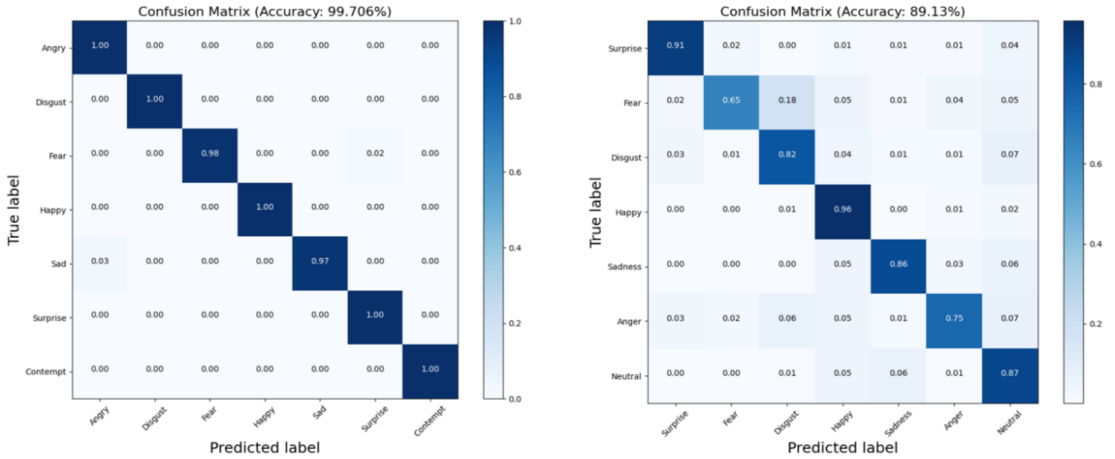

Figure 7: Confusion matrix (Left: CK+, Right: RAF-DB).

To better demonstrate the interpretability of the model, Grad-CAM is utilized to investigate the salient regions of model's attention when performing prediction. From the attention map on the Figure 8, one can see that our proposed network pays more attention to mouth and eyes. Since it's intuitive and natural to focus the subtle difference on these regions, the performance results of our model would work as expected.

5.2. *Ablation study*

To verify the contribution of FDRL and ECA to the proposed model, we conducted ablation experiments. As shown in Table 2, the ECAResNet-18 (w/o FDRL) on the table only got the accuracy of 83.43% on RAF-DB dataset, while it only got 93.43% on CK+ dataset. It is about 5.7% less than the proposed model FDRL-ECANet on RAF-DB dataset and about 6.3% less than the proposed model on CK+ dataset. Even though the results don't show an obvious contribution from ECA alone, still a slight improvement can be observed. On the RAF-DB dataset, the baseline model has an accuracy of 88.47%, which is only 0.65% less than that of our proposed model FDRL-ECANet. On the CK+ dataset, the baseline model has an accuracy of 99.54%, which is only 0.16% less than that of our proposed model FDRL-ECANet.

This ablation study draws two conclusions. First, both ECA and FDRL module can improve the performance for a model to do facial recognition task. Second, both the baseline and ECA alone show no sign of performing better than our proposed model.

Table 2. The result of the ablation study.

Method	CK+ Accuracy (%)	RAF-DB Accuracy (%)
w/o ECA (baseline)	99.54	88.47
w/o FDRL	93.43	83.43
FDRL-ECANet (proposed)	**99.70**	**89.13**

6. Conclusion

In this paper, we proposed a hybrid model of channel attention and feature manipulation to improve the model accuracy of facial expression recognition (FER) by combining an efficient channel attention (ECA) module with a convolutional neural network residual network ResNet-18 and a feature decomposition and reconstruction learning (FDRL) module. We pre-trained the first part of the backbone network ECAResNet-18 on the MS-CelebA dataset and used the ECA module to help better extract the features of the input images from both the CK+ and RAF-DB datasets. The extracted features were then fed into the second part FDRL module to extract intra- and inter-class relationship information and obtain the output results. Our experiment results demonstrate that our model can perform better than the reference model (ResNet-18+FDRL). Both the FDRL and ECA modules can work hard to increase the accuracy of the task of classifying emotions. Our experimental findings demonstrate that the model performs at the cutting edge on both the CK+ and RAF-DB datasets. In the future, our work will more focus on developing novel approaches to better solve the challenge and deploy it in a real-world setting.

References

[1] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q., (2020). ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. arXiv preprint arXiv: 1910. 03151.

[2] He, K., Zhang, X. Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[3] Ruan, D., Yan, Y., Lai, S., Chai, Z., Shen, C., & Wang, H. (2021). Feature decomposition and reconstruction learning for effective facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7660-7669).

[4] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2010, pp. 94-101, doi: 10.1109/CVPRW.2010.5543262.

[5] Li, S., Deng, W., & Du, J. (2017). 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 (pp. 2852-2861).

[6] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

[7] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[8] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

[9] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.

[10] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

[11] Nair, V., & Hinton, G. E. (2010, January). Rectified linear units improve restricted boltzmann machines. In Icml.

[12] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[13] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

[14] Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning (pp. 448-456). PMLR.

[15] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. nature, 323(6088), 533-536.

[16] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. arXiv preprint arXiv:1312.6203, 2013.

[17] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In Proceedings of the European Conference on Computer Vision, pages 486– 504, 2018.

[18] Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In Proceedings of the IEEE international conference on computer vision (pp. 3730-3738).

[19] Xiangyun Zhao,Xiaodan Liang,Luoqi Liu,Teng Li,Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. Peak-piloted deep network for facial expression recognition. In Proceed- ings of the European Conference on Computer Vision, pages 425–442, 2016.

[20] Zibo Meng, Ping Liu, Jie Cai, Shizhong Han, and Yan Tong. Identity-aware convolutional neural network for fa- cial expression recognition. In Proceeding of IEEE Inter- national Conference on Automatic Face & Gesture Recogni- tion), pages 558–565, 2017.

[21] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial ex- pression recognition. IEEE Transactions on Image Process- ing, 28(1):356–370, 2018.

[22] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expres- sion recognition with inconsistently annotated datasets. In Proceedings of the European Conference on Computer Vi- sion, pages 222–237, 2018.

[23] Dinesh Acharya, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Covariance pooling for facial expression recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 367–374, 2018.

[24] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expres- sion recognition by de-expression residue learning. In Pro- ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2168–2177, 2018.

[25] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing, 29:4057–4069, 2020.

[26] Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa. FaceNet2ExpNet: Regularizing a deep face recognition net for expression recognition. In Proceeding of IEEE Interna- tional Conference on Automatic Face & Gesture Recognition, pages 118–126, 2017.

[27] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6897– 6906, 2020.

[28] Delian Ruan, Yan Yan, Si Chen, Jing-Hao Xue, and Hanzi Wang. Deep disturbance-disentangled learning for facial ex- pression recognition. In Proceedings of the 28th ACM In- ternational Conference on Multimedia, pages 2833–2841, 2020.