# SDNET: LIGHTWEIGHT FACIAL EXPRESSION RECOGNITION FOR SAMPLE DISEQUILIBRIUM

*Lifang Zhou*[1,2,3], *Siqin Li*[1,3], *Yi Wang*[2,3,*], *Junlin Liu*[2,3]

[1]College of Computer Science and Technology,
Chongqing University of Posts and Telecommunications, Chongqing, China
[2]College of Software, Chongqing University of Posts and Telecommunications, Chongqing, China
[3]Chongqing Key Laboratory of Image Cognition,
Chongqing University of Posts and Telecommunications, Chongqing, China

## ABSTRACT

Facial expression recognition (FER) based on the convolutional neural network (CNN) in the wild have numerous challenges. For instance, the complexity of the network model makes FER tasks difficult to deploy on portable devices. Some approaches design lightweight networks to reduce the model size, while the intrinsic imbalance of the existing facial emotion datasets is still ignored. In order to overcome the above problems, the lightweight CNN based on sample equalization method for FER is designed to reduce the network parameters sharply while maintaining the identification accuracy. Specifically, to reduce the number of network parameters, a lightweight network framework (SDNet) is designed with separable convolution layers and dense blocks, which can significantly reduce network parameters. Second, the adaptive class weights are proposed to solve the imbalance of sample numbers. Moreover, a resist overfitting (RO) loss function is proposed to improve the classification accuracy. Extensive experiments are conducted on lab-controlled datasets (CK+, Oulu-CASIA) and in-the-wild datasets (FER2013, SFEW). Experimental results show that our method is superior to several state-of-the-art FER methods.

***Index Terms***— Facial expression recognition, lightweight network, sample equalization

## 1. INTRODUCTION

Facial expression is one of the most natural, forceful, and general symbols for human beings to express emotions, which plays an important role in human daily communication. Facial expression recognition (FER) has attracted increasing attention due to its extensive application in man-machine interaction, including human-computer interaction, driver fatigue surveillance, and sociable robotics.

The original FER traditional methods utilize handcrafted features such as neighborhood-aware edge directional pattern (NEDP) [1], local prominent directional pattern (LPDP) [2], and scale-invariant feature transform (SIFT) [3]. But with the advent of in-the-wild datasets like FER2013 [4] and SFEW [5], FER methods based on convolutional neural networks (CNN) have attracted considerable attention because of their high recognition rate. However, deep learning often generates a lot of parameters and cannot meet the requirements of practical applications. In recent years, researchers have begun designing lighter, faster networks. For storage problems, the usual approach is model compression, which is to compress the trained Model so that the network carries fewer network parameters, but it does not take into account the recognition rate. In 2017, MobileNet [6] was proposed by Howard et al., SqueezeNet [7] was proposed by Iandola et al. These methods designed more efficient networks by changing the convolution mode, so as to reduce network parameters without affecting network performance. However, FER using the networks mentioned above directly may not be effective due to the challenges of in-the-wild datasets. To alleviate this problem, we designed a lightweight network framework, SDNet. This network uses deep separable convolution to reduce network parameters generated by convolution operations and uses dense blocks to enhance the transitivity of deep features at different layers, so as to achieve the purpose of lightweight and high efficiency while ensuring accuracy.

However, the uneven distribution of datasets in facial expression samples is still neglected. As a result, FER systems being trained on the above datasets may behave well on dominant emotions, but they behave poorly on the underrepresented ones. Considering the above issue, Ngo et al. [8] proposed the weighted-cluster loss, which simultaneously improves the intra-class compactness and the inter-class separability by learning a class center for each emotion class. Deep multi-task learning (DMTL) [9] has been designed to extract the local information of sample spatial distribution. However, they did not take into account poor image quality. In order to solve the problem mentioned above, the Expression Conditional GAN (ECGAN) [10] has been proposed to generate

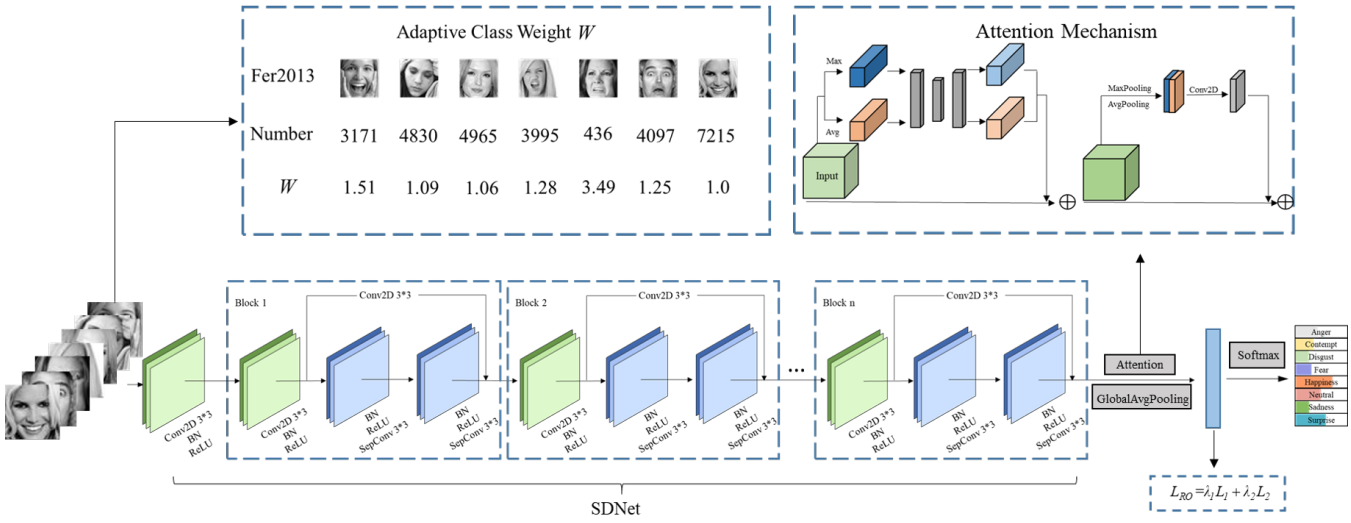


**Fig. 1**. Framework of the proposed facial emotion recognition network.

high-quality facial expression images. However, this method requires the joint execution of generative adversarial networks and convolutional networks, which will generate a large number of parameters, so that recognition speed will sharply decline. Furthermore, Xu et al. [11] proposed local binary convolution (LBC), which was motivated by LBP. Experiments show that the parameters can be significantly reduced. However, the influence of feature extraction can not be ignored. Based on this, we propose an adaptive class weight to act on the loss function. The weight is calculated by the number of samples of each class in the sample, so as to increase the weight of small samples. This can improve the recognition rate of small sample classes while keeping the network light. Moreover, the cross entropy loss function commonly adopted by FER has the risk of overfitting. We design a custom loss function to alleviate this problem.

## 2. PROPOSED METHOD

The proposed facial expression recognition process is shown in Figure 1. It is worth noting that we use the proposed SDNet as the backbone network, which adopts the separable convolution layers and dense blocks. This network can effectively reduce parameter size. We also design the adaptive class weight and the RO loss function to improve the recognition accuracy. Moreover, the convolutional block attention module (CBAM) is also adopted.

### 2.1. Design of Lightweight Network Structure

An applicable FER task requires a lightweight and efficient convolutional network to implement the effects of loss.

We are inspired by XceptionNet [12] and DenseNet to design the network lighter and more efficient. In particular, the deep separable convolution is utilized to reduce the network parameters generated in convolution operation, and the adoption of residual structure for network feature mapping can alleviate the problem of network degradation. Furthermore, the dense blocks are used to enhance the transitivity of deep features in different layers. Based on this, we design a lightweight network framework SDNet with the structure shown in Fig. 1.

### 2.2. Adaptive Class Weight

In the FER task, the difference of sample number will lead to the decrease of average recognition accuracy, so it is necessary to reduce the influence of small sample classes on network training. We propose an adaptive class weight method to increase the weight of small samples, so as to improve the recognition rate of small samples, and finally improve the overall recognition rate.

The first step is to calculate the total number of samples in the expression database and set a super parameter $\partial$ as the weight strength factor. The formula is as follows:

$$if\ S_i < \mathrm{Avg}\left(\sum_{j=1}^{n} S_j\right)$$

$$\partial = \frac{1}{m} \cdot \sum_{i=1}^{n} \frac{\mathrm{Avg}\left(\sum_{j=1}^{n} S_j\right) - S_i}{\mathrm{Avg}\left(\sum_{j=1}^{n} S_j\right)} \tag{1}$$

where $n$ is the number of categories for expression database, $i = [1, 2, \cdots, n]$ , $S_{i/j}$ is the number of samples of the class $i/j$, $m$ is the total number of classes with less than the average sample size. The weight factor $\partial$ is the different degree

between the small sample and the average sample, the sample weight of each class is obtained by the following formula:

$$\mathrm{W}_i = \log\left(\frac{\partial \cdot \sum_{i=1}^{n}(S_i)}{S_i}\right) \tag{2}$$

according to Equation (2), the corresponding weight of each class can be obtained, and then the log value of the class with fewer samples is larger, while the weight value of the class with multiple samples basically remains unchanged. In order to make the sample loss maintain the balance of the class with multiple samples when paying attention to the class with fewer samples, the value of $W_i$ less than 1 is set as 1,

$$W = \begin{cases} 1.0 & W_i <= 1.0 \\ W_i & W_i > 1.0 \end{cases} \tag{3}$$

the resulting class weight $W$ is used to weight the loss function to balance the quantitative differences between classes.

### 2.3. RO Loss Function

In the classification tasks, most methods often use softmax function for output and cross entropy as loss function. However, the above approach usually leads to the risk of overfitting. Therefore, we design a loss function to solve this problem, the formula is as follows:

$$L_{RO} = L_1 + L_2 \tag{4}$$

$$L_1 = \lambda_1 \sum_{i=1}^{n}\left(p\left(x_i\right)\log\left(q\left(x_i\right)\right)\right) \tag{5}$$

$$L_2 = \lambda_2 \sum_{i=1}^{n}\left[\frac{O\left(p\left(x_i\right)\right)}{N}\log\left(q\left(x_i\right)\right)\right] \tag{6}$$

where $\lambda_1$ , $\lambda_2$ is the hyperparameter factor, $p\left(x_i\right)$ is the label value, and $q\left(x_i\right)$ is the predicted value, $O\left(p\left(x_i\right)\right)$ creates a tensor with the same dimension and elements of 1 as the label value, $N$ is the number of classes.

## 3. EXPERIMENTS AND PERFORMANCE

In this section, we introduce the detailed experimental setup, including the datasets, comparative experiments, and experimental analysis.

### 3.1. Datasets

We mainly use two types of datasets, including lab-controlled datasets (The Extended Coh-Kanade dataset (CK+) [16], Oulu-CASIA) [17]) and in-the-wild datasets (FER2013, SFEW).

The CK+ contains 593 image sequences from 123 subjects, the last frame of each image sequence is labeled with the action unit. Of all the image sequences, 327 were labeled with emotion. We choose the last three frames of the image sequence as the peak frames and take the first frame of each sequence as the neutral expression, then the dataset can be divided into 8 types of expressions [18]. There are no neutral and contemptuous expressions when CK+ is divided into 6 categories, and no neutral expressions when it is divided into 7 categories.

The Oulu-CASIA dataset uses 480 videos collected by the VIS system under normal indoor lighting. For each video sequence, the last three frames are taken as the peak frames of expression, and the dataset contains 1440 images [19].

The FER2013 dataset is a large public dataset, which consists of 35,886 facial expression images, including 28,709 training images, 3,589 verification images, and 3,589 test images. The dataset contains seven expressions including anger, disgust, fear, happiness, sadness, surprise, and neutral.

The SFEW dataset has 879 training samples and 406 validation samples, which are collected from movies.

### 3.2. Experiment Settings

The input image size of the model is set as $68 \times 68$, the weight decay is 0.0004, the learning rate is initialized to 0.001, and the batch size is 16. The GPU of the experimental server is NVIDIA GTX1050, and the memory is 8.0GB. The model is implemented using KERAS.

### 3.3. Experiment Results

#### *3.3.1. Comparison with the lightweight network model*

In order to prove the lightness and superiority of SDNet proposed in this paper, we compared SDNet with other lightweight networks in terms of parameter size, FLOPs, and accuracy in different datasets. The adaptive weights proposed in this paper and the improved losses are uniformly applied in the experiment. The experimental results are shown in Table 1.

Experimental results show that our SDNet is only 0.63m bigger than MobileNetV3 in parameters and 2.29m bigger than it in FLOPs. But our recognition rate is much better than was much higher than MobileNetV3's, improved 30.89% on CK+, 28.13% on Oulu-CASIA, and 22.37% on SFEW.

In terms of network parameter size, our proposed network is only one-tenth of XceptionNet. FLOPS are a third of that. It is worth noting that our network is higher in accuracy. This shows that SDNet has a small number of parameters and FLOPs while maintaining a high accuracy, which can be well deployed in mobile applications.

#### *3.3.2. Comparison with State-of-the-Art*

The performance of the proposed method was compared with state-of-the-art methods on CK+, Oulu-CASIA, FER2013,

**Table 1**. Comparison of different lightweight network models

| Model | CK+ | Oulu-CASIA | FER2013 | SFEW | Parameters | FLOPs |
|---|---|---|---|---|---|---|
| MobileNetV3 [13] | 66.90% | 68.75% | 64.78% | 23.29% | **1.54M** | **3.09M** |
| ShuffleNetV2 [14] | 77.94% | 85.42% | 62.45% | 42.01% | 2.30M | 5.72M |
| GghostNet [15] | 91.18% | 92.01% | **67.73%** | 45.29% | 3.27M | 6.58M |
| XceptionNet [12] | **97.79%** | 96.53% | 66.52% | 41.55% | 20.87M | 17.2M |
| **Ours(SDNet)** | **97.79%** | **96.88%** | 67.08% | **45.66%** | 2.17M | 5.38M |

**Table 2**. Comparison of FER results of different methods on CK+

| Method | Accuracy |
|---|---|
| LDL-ALSG[20] | 93.08% |
| IPA2LT [21] | 91.67% |
| AGRA [22] | 85.27% |
| Li et al. [23] | 96.46% |
| DAM-CNN [24] | 95.88% |
| **Ours** | **97.79%** |

**Table 3**. Comparison of FER results of different methods on Oulu-CASIA

| Method | Accuracy |
|---|---|
| LDL-ALSG[20] | 63.94% |
| IPA2LT [21] | 61.02% |
| Kuo et al. [25] | 88.75% |
| STC-NLSTM [26] | 93.45% |
| FESGAN [27] | 88.13% |
| **Ours** | **96.88%** |

**Table 4**. Comparison of FER results of different methods on FER2013

| Method | Accuracy |
|---|---|
| Li et al. [23] | 54.81% |
| AGRA [22] | 58.95% |
| DAM-CNN [24] | 65.31% |
| **Ours** | **67.08%** |

**Table 5**. Comparison of FER results of different methods on SFEW

| Method | Accuracy |
|---|---|
| CycleAT [28] | 30.75% |
| Sun et al. [29] | 40.00% |
| DAM-CNN [24] | 42.30% |
| **Ours** | **45.66%** |

and SFEW databases respectively. The experimental results are shown in Table 2-5.

The experimental results in the above table show that the proposed method performs better than the existing methods on lab-controlled datasets and in-the-wild datasets. Specifically, in terms of accuracy, our method improved by 1.33% compared with Li et al. and 12.52% compared with AGRA on CK+. On Oulu-CASIA, the recognition rate increased from 63.94% to 96.88% compared with LDL-ALSG. Compared with DAM-CNN, the accuracy rate increased from 65.31% to 67.08% on FER2013 and from 42.30% to 45.66% on SFEW.

## 4. CONCLUSIONS

This paper proposes a Lightweight CNN facial expression Recognition based on sample equalization. The designed SDNet combines the advantages of the separable convolution layer and the dense block to dramatically reduce parameters and improve efficiency. We also propose an adaptive class weight act on the loss function to alleviate sample imbalance. Moreover, We also design the RO loss function to avoid overfitting. The proposed method is evaluated on lab-controlled datasets and in-the-wild datasets, experimental results show that this method performs well in both lightweight and accuracy.

## 5. ACKNOWLEDGEMENTS

This work was supported by the Science and Technology Research Program of Chongqing Graduate Scientific Research Innovation Project (CYS21323).

# References

[1] M.T. Iqbal, M Abdullah-Al-Wadud, and B.Y. Ryu, "Facial expression recognition with neighborhood-aware edge directional pattern (nedp)," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 125–137, 2018.

[2] F. Makhmudkhujaev, M. Abdullah-Al-Wadud, and M.T. Iqbal, "Facial expression recognition with local prominent directional pattern," *Signal Processing: Image Communication*, vol. 74, pp. 1–12, 2019.

[3] F.J. Ren and Z. Huang, "Facial expression recognition based on aam–sift and adaptive regional weighting," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 10, no. 6, pp. 713–722, 2015.

[4] I.J. Goodfellow, D. Erhan, and P.L. Carrie, "Challenges

in representation learning: A report on three machine learning contests," in *International conference on neural information processing*. Springer, 2013, pp. 117–124.

[5] A. Dhall, R. Goecke, and S. Lucey, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2011, pp. 2106–2112.

[6] A.G. Howard, M.L. Zhu, and B. Chen, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[7] F.N. Iandola, S. Han, and M.W. Moskewicz, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[8] Q.T. Ngo and S. Yoon, "Facial expression recognition based on weighted-cluster loss and deep transfer learning using a highly imbalanced dataset," *Sensors*, vol. 20, no. 9, pp. 2639, 2020.

[9] H.Zheng, R.L. Wang, and W.T.Ji, "Discriminative deep multi-task learning for facial expression recognition," *Information Sciences*, vol. 533, pp. 60–71, 2020.

[10] H. Tang, W. Wang, and S.S Wu, "Expression conditional gan for facial expression-to-expression translation," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 4449–4453.

[11] J.F. Xu, Felix, and B. Naresh, "Local binary convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 19–28.

[12] C. François, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[13] A. Howard, M. Sandler, and G. Chu, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1314–1324.

[14] N.N. Ma, X.Y. Zhang, and H.T. Zheng, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[15] K. Han, Y.H. Wang, and Q. Tian, "Ghostnet: More features from cheap operations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1580–1589.

[16] 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 *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2010, pp. 94–101.

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

[18] J. Cai, Z.B. Meng, and A. S. Khan, "Island loss for learning discriminative features in facial expression recognition," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 302–309.

[19] I. Cugu, E. Sener, and E. Akbas, "Microexpnet: An extremely small and fast model for expression recognition from face images," in *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 2019, pp. 1–6.

[20] S.K. Chen, J.F. Wang, and Y.D. Chen, "Label distribution learning on auxiliary label space graphs for facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13984–13993.

[21] J.B. Zeng, S.G. Shan, and X.L. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 222–237.

[22] Y. Xie, T.S. Chen, and T. Pu, "Adversarial graph representation adaptation for cross-domain facial expression recognition," in *Proceedings of the 28th ACM international conference on Multimedia*, 2020, pp. 1255–1264.

[23] W.H. Li, S.and Deng, "Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning," *International Journal of Computer Vision*, vol. 127, no. 6, pp. 884–906, 2019.

[24] S.Y. Xie, H.F. Hu, and Y.B. Wu, "Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition," *Pattern recognition*, vol. 92, pp. 177–191, 2019.

[25] C.M. Kuo, S.H. Lai, and M. Sarkis, "A compact deep learning model for robust facial expression recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 2121–2129.

[26] Z.B. Yu, G.C. Liu, and Q.S. Liu, "Spatio-temporal convolutional features with nested lstm for facial expression recognition," *Neurocomputing*, vol. 317, pp. 50–57, 2018.

[27] Y. Yan, Y. Huang, and S. Chen, "Joint deep learning of facial expression synthesis and recognition," *IEEE Transactions on Multimedia*, vol. 22, no. 11, pp. 2792–2807, 2019.

[28] F.F. Zhang, T.Z. Zhang, and Q.R. Mao, "Facial expression recognition in the wild: A cycle-consistent adversarial attention transfer approach," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 126–135.

[29] W.Y. Sun, H.T. Zhao, and Z. Jin, "A visual attention based roi detection method for facial expression recognition," *Neurocomputing*, vol. 296, pp. 12–22, 2018.