MoVE-CNNs: Model aVeraging Ensemble of Convolutional Neural Networks for Facial Expression Recognition

Jing Xuan Yu, Kian Ming Lim, *Member, IAENG,* and Chin Poo Lee, *Member, IAENG*

***Abstract*—Facial expression is a powerful non-verbal communication that can express emotions and messages without saying a single word. In view of the prominence of facial expression, we propose a model averaging ensemble of Convolutional Neural Networks (CNN) that consolidates multiple pre-trained CNN models. Each pre-trained CNN model first undergoes transfer learning with the classification layer substituted with a multilayer perceptron. The newly formed model is then fine-tuned on the facial expression datasets and adapted to facial expression recognition. The predictions returned by all models are combined by model averaging to determine the final class probability distributions. The proposed model averaging ensemble of CNNs is evaluated on three facial expression datasets: FER-2013, modified CK+ and RAF-DB. Since the modified CK+ dataset is a small dataset, data augmentation is leveraged to increase the size and diversity of data. Apart from that, oversampling is adopted to address the class imbalance challenge in RAF-DB. The empirical results demonstrate that the proposed model averaging ensemble of CNNs outperforms the individual ensemble model at the test accuracy of 77.70%, 94.10% and 87.50% in FER 2013, modified CK+ and RAF-DB datasets, respectively.**

***Index Terms*—facial expression, facial expression recognition, convolutional neural network, ensemble, model averaging, transfer learning, data augmentation, oversampling**

I. Introduction

NON-VERBAL communication is a means of communication without the use of auditory language to convey the message, with facial expression being one of them. The high expressivity of facial expression enables people to convey countless emotions without uttering a single word. There are six basic emotions that are identified as the universal emotions, namely anger, disgust, fear, happiness, sadness and surprise [1]. Later, contempt was added to the list of universal emotions [2].

The rise of artificial intelligence has transformed the way people work and greatly improves the ability of machines in complicated tasks. The earlier works on facial expression recognition were dominated by hand-engineered methods for feature extraction [3], [4], [5]. Hand-engineered methods refer to the non-learning-based methods where the feature representations are manually crafted and then used for classification [6], [7]. In recent years, learning-based methods are widely adopted in facial expression recognition. Learning-based methods automatically learn the discriminative features from the input and subsequently use the learned features for classification [8], [9]. Deep learning is a subset of learning-based methods that has shown superior performance in a wide spectrum of applications, such as computer vision, natural language processing, speech processing, signal processing, and so like. In the field of computer vision, Convolutional Neural Networks (CNN) are amongst the most influential and popular deep learning methods.

In this paper, an ensemble method that integrates several CNN models, referred to as "Model aVeraging Ensemble of CNNs (MoVE-CNNs)" is leveraged for facial expression recognition. The CNN models include VGG16 pre-trained on VGG-Face dataset, VGG16, VGG19, ResNet50 and ResNet101 pre-trained on ImageNet. To enhance the performance of each model, two enhancements namely transfer learning and model averaging ensemble are proposed. Transfer learning is performed on each model where the last few layers of the pre-trained model are replaced with a multilayer perceptron (MLP). Subsequently, these CNN models are combined using model averaging ensemble method to aggregate the class probability distributions returned by each model.

The performance of MoVE-CNNs is evaluated on three datasets: Facial Expression Recognition 2013 (FER-2013) dataset, modified Extended Cohn-Kanade (CK+) dataset and Real-world Affective Face Database (RAF-DB). The modified CK+ is a subset of CK+ dataset that only contains 981 face images from the frontal view. Since the number of face images is relatively small and insufficient to train a decent CNN model, data augmentation is applied on the modified CK+ dataset. Data augmentation generates transformed variations of the training images to increase the training data size and to address the data scarcity issue. Apart from that, RAF-DB consists of images from seven basic emotions but the sample size of each emotion is imbalanced. Therefore, oversampling is implemented on RAF-DB to adjust the class distribution by creating new images for the minority class. The empirical results demonstrate that MoVE-CNNs outshines the individual contributing CNN model and enhances the overall performance in facial expression recognition. To this end, the main contributions of this paper are:

- Transfer learning on five pre-trained CNN models where a multilayer perceptron is concatenated to each model. The multilayer perceptron consists of a global pooling layer, fully-connected layers and classification layer.

Manuscript received May 17, 2021; revised August 5, 2021. This work was supported by Fundamental Research Grant Scheme of the Ministry of Higher Education under award number FRGS/1/2019/ICT02/MMU/03/7.

J.X. Yu is a student in the Faculty of Information Science and Technology, Multimedia University, Melaka, 75450 Malaysia. (e-mail: jingxuan97717@gmail.com)

K.M. Lim, the corresponding author, is a Lecturer in the Faculty of Information Science and Technology, Multimedia University, Melaka, 75450 Malaysia. (phone: 606-2523066; e-mail: kmlim@mmu.edu.my)

C.P. Lee is a Senior Lecturer in the Faculty of Information Science and Technology, Multimedia University, Melaka, 75450 Malaysia. (e-mail: cplee@mmu.edu.my)

The concatenated models are thereafter fine-tuned on the facial expression datasets.
- A model averaging ensemble of CNNs (MoVE-CNNs) to combine the predictions of the fine-tuned models to improve the overall performance in facial expression recognition. The model averaging ensemble outshines the average performance of any single ensemble member.
- Data augmentation is performed on the modified CK+ dataset to generate transformations of the images hence increases the data size by 5 times. A larger data size enhances the generalization capability of the trained model and leads to more robust facial expression recognition.
- The SMOTE technique oversamples the minority class in RAF-DB. The oversampling ameliorates the class balance and alleviates the bias caused by the skewed dataset.

II. Related Works

This section describes some state-of-the-art facial expression recognition using CNN and its variants, transfer learning and ensemble learning.

In [10], a modified CNN was proposed by applying batch normalization to the output of the first and the last convolutional layers. Batch normalization ensures the input to the subsequent layers are normalized according to the mean and standard deviation of the batch, thus stabilizing and expediting the training process. A shallow CNN model for facial expression recognition was proposed in [11] to alleviate the requirement of large training data. The authors in [12] used some pre-processing such as cropping the face region, histogram equalization and downsampling before passing into CNN. To increase the training data size, random rotation and flipping were applied to the training set.

In [13], two subject-specific facial expression models were proposed via inductive and transductive transfer learning. Using inductive transfer learning, the model can be propagated to a new subject only with a small number of subspecific training data. On the other hand, transductive transfer learning eliminates the need of data labeling. In [14], a facial expression recognition model with transfer learning of a CNN model was proposed. The model was tested in an occluded condition and showed promising performance in the small occlusion case. The authors of [15] presented a cascaded fine-tuning where the CNN pre-trained on ImageNet was fine-tuned on two facial expression datasets sequentially. They suggested cascaded fine-tuning performs better than one-time fine-tuning on combined datasets. The work [16] compared the performance of AlexNet and VGG16 with transfer learning and end-to-end training. They suggested that transfer learning performs better and requires shorter training time than end-to-end training.

The work [17] learned several weak classifiers for each facial expression using the source dataset. Subsequently, a strong classifier is built for each facial expression by combining the weak classifiers. An ensemble model of three face detectors were leveraged in [18] to detect and extract faces from Static Facial Expressions in the Wild (SFEW) dataset. Subsequently, the classification was done by an ensemble of multiple CNNs. The authors in [19] presented a CNN model with an ensemble of three subnets. These subnets were trained separately and then assembled together by replacing the output layers with a fully connected layer at the end.

III. Model aVeraging Ensemble of CNNs (MoVE-CNNs)

In this work, we propose an ensemble method that integrates multiple CNN models, referred to as Model aVeraging Ensemble of CNNs (MoVE-CNNs). The MoVE-CNNs method adopts five CNN models, namely VGG16 [20] pre-trained on VGG-Face dataset [21] (denoted as "VGG16-1" in this paper) and four CNN models pre-trained on ImageNet, namely VGG16 (denoted as "VGG16-2" in this paper), VGG19, ResNet50 [22] and ResNet101. Firstly, transfer learning is performed on each pre-trained model to fine-tune the model specifically for facial expression dataset. Subsequently, these fine-tuned models work as an ensemble in facial expression recognition. The class probability distributions returned by the ensemble model are compiled by the model averaging method.

A. Transfer Learning

Transfer learning is an idea of machine learning where a model built for one task is fine-tuned for another new task. Using transfer learning, the representative features learnt from the source task are propagated to the new target task with minimal re-training. The intuition is that, since the model is already pre-trained on a large dataset (e.g. ImageNet), a good set of discriminative features is captured thus can be optimally used to describe the data specific to the new task. A key benefit of transfer learning is to circumvent the need of training the model from scratch, hence greatly reduces the size of training dataset and training time.

In this work, five pre-trained models, namely VGG16-1, VGG16-2, VGG19, ResNet50 and ResNet101 are used as the feature extractor of the new model. This is done by loading the pre-trained models without the final classification layer. Thereafter, a multilayer perceptron (MLP) comprises a global maxpooling layer, fully-connected layers and classification layer is concatenated to the pre-trained model. The newly formed models are then re-trained on FER-2013, modified CK+ and RAF-DB dataset separately. The layers of the pre-trained models are frozen and the re-training only fine-tunes the layers of the MLP.

Unlike pooling layer that downsampled the patches of the feature map, a global pooling layer downsamples the entire feature map to a single value. The global maxpooling summarizes a feature map spatially by extracting the maximum activation in the feature map. Global maxpooling is efficient in denoising and dimension reduction which leads to better translational invariance and less overfitting in the model. The output from the global maxpooling layer is then flattened and passed into the fully-connected layer. The fully-connected layer learns the discriminative features that are specific to the facial expression dataset by updating its weights to maximize the accuracy and minimize the function loss. Lastly, the facial expression recognition is performed in the classification layer where a softmax function is leveraged to compute the probability of each class.

B. Model Averaging Ensemble

Ensemble learning is the process where a myriad of machine learning models are combined to perform a particular task. The models that contribute to the ensemble are referred to as ensemble members. The ensemble learning combines the decisions from the ensemble members to enhance the overall classification performance. In this work, a simple ensemble learning technique, known as model averaging ensemble is leveraged. A simple ensemble learning technique is more computationally inexpensive compared to advanced ensemble learning techniques such as boosting and bagging. The key advantages of ensemble learning are to reduce the variance of the predictions and to improve the average performance of the ensemble members.

Here, the ensemble members are the fine-tuned VGG16-1, VGG16-2, VGG19, ResNet50 and ResNet101 with MLP after transfer learning. The architecture of the model averaging ensemble is illustrated in Figure 1. Each model makes their predictions and returns the probability distributions of the facial expression classes. The probability distributions are thereafter combined by the averaging procedure, defined as:

$$\tilde{y}(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{n} y_j(\mathbf{x}) \tag{1}$$

where $\tilde{y}(\mathbf{x})$ denotes the average probability of a facial expression class $\mathbf{x}$, $y_j(\mathbf{x})$ represents the probability of a facial expression class returned by the j-th ensemble member and n represents the number of ensemble members.

IV. Dataset

The MoVE-CNNs method is evaluated on three facial expression datasets, including Facial Expression Recognition 2013 (FER-2013) dataset, modified Extended Cohn-Kanade (CK+) dataset and Real-world Affective Face Database (RAF-DB).

The FER-2013 dataset [23] contains 28709 train data, 3589 validation data and 3589 test data. The images are categorized to one of the seven expressions, namely happiness, surprise, neutral, anger, disgust, fear and sadness. The images in FER-2013 are normalized into grayscale images with 48×48 pixels.

Extended Cohn-Kanade (CK+) dataset [24] comprises 593 video sequences of 123 subjects. Each sequence shows a conversion from calm emotion to extreme emotion. Out of 593 videos, 327 sequences of 118 subjects are labelled with one of the emotion labels. Since frontalization is not the focus of this work, we adopt a subset of CK+ dataset that only consists of 981 facial expression images from the frontal view. The subset is referred to as the modified CK+ dataset.

Real-world Affective Face Database (RAF-DB) [25], [26] contains approximately 30000 real-world face images downloaded from the Internet. The samples are labelled into seven basic emotions and twelve compound emotions. In this work, only 15339 images of basic emotions are used in the experiments.

A. Data Augmentation

Since the modified CK+ dataset only contains 981 images, data augmentation is performed to alleviate the data scarcity issue. Data augmentation increases the data size by constructing the transformed versions of the original images in the dataset. By providing more diversified images for training, the model can learn more discriminative features and be less susceptible to overfitting. The data augmentation techniques used in this work include height shift, width shift, rotation and horizontal flipping. By applying data augmentation on the modified CK+ dataset, the data size is increased by a factor of 5.

B. Oversampling

The RAF-DB dataset suffers from the class imbalance issue where some facial expressions have much larger data size than others. To this end, oversampling is applied mitigate the class imbalance issue. The oversampling technique increases the data samples of the minority class by creating new instances of the class. This paper leverages Synthetic Minority Over-sampling Technique (SMOTE) [27] for oversampling. In SMOTE, the entire dataset is used as the input, a sample of the minority class is then randomly selected and the k-nearest neighbours of the selected sample are identified. Based on the k-nearest neighbours, new synthetic samples are generated. In doing so, the class imbalance issue is alleviated.

V. Experiments and Discussions

In the experiments, all datasets are partitioned into 60% training, 20% validation and 20% test set. Data augmentation is performed on the modified CK+ dataset to enlarge the sample size, whilst SMOTE is applied on RAF-DB to ameliorate the balance in the class size. The experimental results of MoVE-CNNs and its ensemble members are presented in Table I.

Table I
The experimental results of MoVE-CNNs and its ensemble members on FER-2013, modified CK+ and RAF-DB.

Model	**Accuracy (%)**		
	FER-2013	**Modified CK+**	**RAF-DB**
Fine-tuned VGG16-1 with MLP	53.55	88.32	70.35
Fine-tuned VGG16-2 with MLP	50.15	92.39	63.57
Fine-tuned VGG19 with MLP	54.95	90.86	71.02
Fine-tuned ResNet50 with MLP	68.96	75.63	65.97
Fine-tuned ResNet101 with MLP	77.51	79.19	71.72
MoVE-CNNs	**77.70**	**94.10**	**87.50**

For the FER-2013 dataset, the MoVE-CNNs method greatly improves the accuracy of the fine-tuned VGG16-1, VGG16-2, VGG19 and ResNet50 by around 8-24%. The fine-tuned ResNet101 performs comparatively with MoVE-CNNs in FER-2013. As for the modified CK+ dataset, the MoVE-CNNs method achieves an accuracy of 94.10% despite the small data size. It is attributed to the transfer learning that enhances the feature representation and the data augmentation that produces more training samples for better generalization capability of the model. For the modified CK+ dataset, the MoVE-CNNs method outperforms the individual ensemble members by approximately 2-19%.

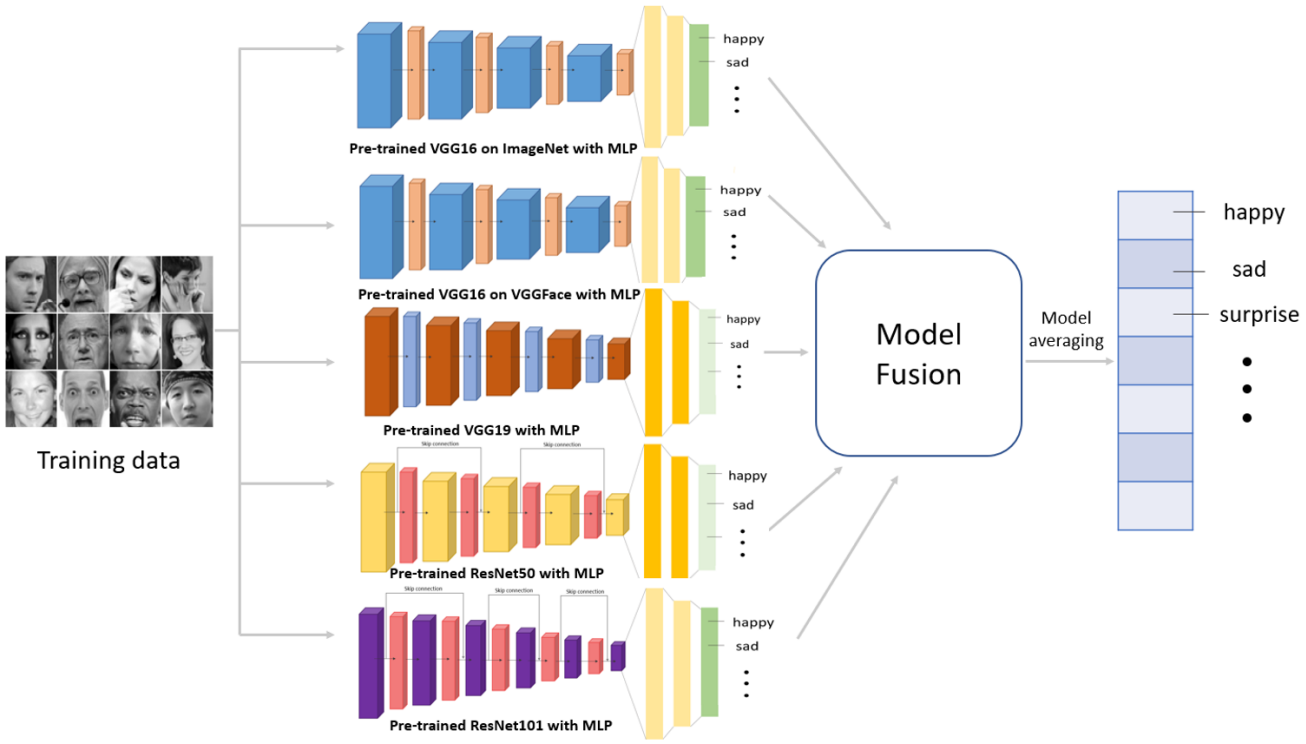

Figure 1. The architecture of the model averaging ensemble of fine-tuned VGG16-1, VGG16-2, VGG19, ResNet50 and ResNet101 with MLP.

Similar trends are observed in RAF-DB where the MoVE-CNNs method yields a test accuracy of 87.5%, which is 16-24% improvements over the individual ensemble members. The promising results on RAF-DB are also ascribable to the SMOTE technique that mitigates the imbalance in the class size. The empirical results corroborate the MoVE-CNNs method that outshines the individual ensemble member and is effective in improving the overall performance by combining the predictions of the contributing members.

VI. Conclusion

Facial expression is a powerful communication medium that can convey countless messages and emotions without uttering a single word. In this paper, we present a model averaging ensemble of CNNs (MoVE-CNNs) method for facial expression recognition. The MoVE-CNNs method combines some pre-trained CNN models, namely VGG16 trained on VGG-Face dataset, VGG16, VGG19, ResNet50 and ResNet101 trained on ImageNet. The pre-trained CNN models mainly function as the feature extractor to capture the discriminative features in the dataset. The classification layer of the pre-trained model is replaced with a multi-layer perceptron consisting of a global pooling layer, fully-connected layers and classification layers. Transfer learning is then performed on the newly formed pre-trained models where they are fine-tuned on the facial expression datasets to better adapt to the new task. The fine-tuned models are then consolidated as a model averaging ensemble where the predictions of all models are averaged to obtain the final probability distributions of the facial expression class.

The proposed MoVE-CNNs method and its ensemble members are evaluated on FER-2013, modified CK+ and RAF-DB datasets. Since the modified CK+ dataset only contains 981 images, data augmentation is applied to increase the sample size by producing transformed variations of the images. Besides that, SMOTE oversampling is leveraged on the RAF-DB dataset to alleviate the class imbalance issue. The SMOTE oversampling creates new instances of the minority class based on the k-nearest neighbours of the samples in the class. The empirical results demonstrate that the proposed MoVE-CNNs method outperforms all ensemble members and effectively improves the overall performance in facial expression recognition. Despite the small number of data samples in the modified CK+ dataset, the model yields a test accuracy of 94.10% attributed to the data augmentation that provides more data samples for model learning. The MoVE-CNNs method achieves a promising accuracy of 87.50% in RAF-DB showing that the oversampling technique is able to ameliorate the class balance.

References

[1] P. Ekmann, "Universal facial expressions in emotion," *Studia Psychologica*, vol. 15, no. 2, p. 140, 1973.

[2] D. Matsumoto, "More evidence for the universality of a contempt expression," *Motivation and Emotion*, vol. 16, no. 4, pp. 363–368, 1992.

[3] S. Sumpeno, M. Hariadi, and M. H. Purnomo, "Facial emotional expressions of life-like character based on text classifier and fuzzy logic," *IAENG International Journal of Computer Science*, vol. 2, no. 38, pp. 122–133, 2011.

[4] Y. Kristian, H. Takahashi, E. Purnama, I. Ketut, K. Yoshimoto, E. I. Setiawan, E. Hanindito, and M. H. Purnomo, "A novel approach on infant facial pain classification using multi stage classifier and geometrical-textural features combination." *IAENG International Journal of Computer Science*, vol. 44, no. 1, pp. 112–121, 2017.

[5] M. Monwar, S. Rezaei, and K. Prkachin, "Eigenimage based pain expression recognition," *IAENG International Journal of Applied Mathematics*, vol. 36, no. 2, pp. 1–6, 2007.

[6] C. P. Lee, A. Tan, and K. Lim, "Review on vision-based gait recognition: Representations, classification schemes and datasets," *American Journal of Applied Sciences*, vol. 14, no. 2, pp. 252–266, 2017.

[7] K. M. Lim, A. W. Tan, and S. C. Tan, "A four dukkha state-space model for hand tracking," *Neurocomputing*, vol. 267, pp. 311–319, 2017.

[8] Y. S. Tan, K. M. Lim, C. Tee, C. P. Lee, and C. Y. Low, "Convolutional neural network with spatial pyramid pooling for hand gesture recognition," *Neural Computing and Applications*, vol. 33, no. 10, pp. 5339–5351, 2021.

[9] Y. S. Tan, K. M. Lim, and C. P. Lee, "Hand gesture recognition via enhanced densely connected convolutional neural network," *Expert Systems with Applications*, vol. 175, p. 114797, 2021.

[10] M. Coşkun, A. Uçar, Ö. Yildirim, and Y. Demir, "Face recognition based on convolutional neural network," in *2017 International Conference on Modern Electrical and Energy Systems (MEES)*. IEEE, 2017, pp. 376–379.

[11] S. Miao, H. Xu, Z. Han, and Y. Zhu, "Recognizing facial expressions using a shallow convolutional neural network," *IEEE Access*, vol. 7, pp. 78 000–78 011, 2019.

[12] K. Li, Y. Jin, M. W. Akram, R. Han, and J. Chen, "Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy," *The visual computer*, vol. 36, no. 2, pp. 391–404, 2020.

[13] J. Chen, X. Liu, P. Tu, and A. Aragones, "Person-specific expression recognition with transfer learning," in *2012 19th IEEE International Conference on Image Processing*. IEEE, 2012, pp. 2621–2624.

[14] M. Xu, W. Cheng, Q. Zhao, L. Ma, and F. Xu, "Facial expression recognition based on transfer learning from deep convolutional networks," in *2015 11th International Conference on Natural Computation (ICNC)*. IEEE, 2015, pp. 702–708.

[15] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 443–449.

[16] I. Oztel, G. Yolcu, and C. Oz, "Performance comparison of transfer learning and training from scratch approaches for deep facial expression recognition," in *2019 4th International Conference on Computer Science and Engineering (UBMK)*. IEEE, 2019, pp. 1–6.

[17] R. Zhu, T. Zhang, Q. Zhao, and Z. Wu, "A transfer learning approach to cross-database facial expression recognition," in *2015 International Conference on Biometrics (ICB)*. IEEE, 2015, pp. 293–298.

[18] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 435–442.

[19] K. Liu, M. Zhang, and Z. Pan, "Facial expression recognition with cnn ensemble," in *2016 international conference on cyberworlds (CW)*. IEEE, 2016, pp. 163–166.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[21] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," 2015.

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

[23] P. Carrier, A. Courville, I. Goodfellow, M. Mirza, and Y. Bengio, "Fer-2013 face database, université de montréal," 2013.

[24] 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," in *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE, 2010, pp. 94–101.

[25] S. Li, W. Deng, and J. Du, "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*, 2017, pp. 2852–2861.

[26] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2018.

[27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.