ResEmoteNet: Bridging Accuracy and Loss Reduction in Facial Emotion Recognition

Arnab Kumar Roy, Hemant Kumar Kathania, Adhitiya Sharma, Abhishek Dey and Md. Sarfaraj Alam Ansari

***Abstract*—The human face is a silent communicator, expressing emotions and thoughts through it's facial expressions. With the advancements in computer vision in recent years, facial emotion recognition technology has made significant strides, enabling machines to decode the intricacies of facial cues. In this work, we propose ResEmoteNet, a novel deep learning architecture for facial emotion recognition designed with the combination of Convolutional, Squeeze-Excitation (SE) and Residual Networks. The inclusion of SE block selectively focuses on the important features of the human face, enhances the feature representation and suppresses the less relevant ones. This helps in reducing the loss and enhancing the overall model performance. We also integrate the SE block with three residual blocks that help in learning more complex representation of the data through deeper layers. We evaluated ResEmoteNet on three open-source databases: FER2013, RAF-DB, and AffectNet, achieving accuracies of 79.79%, 94.76%, and 72.39%, respectively. The proposed network outperforms state-of-the-art models across all three databases. The source code for ResEmoteNet is available at https://github.com/ArnabKumarRoy02/ResEmoteNet.**

***Index Terms*—Facial Emotion Recognition, Convolutional Neural Network, Squeeze and Excitation Network, Residual Network.**

I. Introduction

Facial Emotion Recognition (FER) is a specialized task within Image Recognition, focusing on identifying emotions from facial images or videos. Facial emotions change with subtle movements of facial features such as lips, teeth, skin, hair, cheekbones, nose, face shape, eyebrows, eyes, jawline, and mouth, making it difficult to design models that can accurately capture these intricate details. Additionally, data collection for FER is a labor-intensive process that requires significant funding and careful annotation by humans. Despite its challenges, FER is a crucial task in image recognition. Facial emotions provide valuable insights into a person's mental health, helping to identify signs of depression, anxiety, and other psychiatric disorders [1]. Consequently, FER plays a significant role in mental health and therapy. It can also be instrumental in creating responsive and adaptive systems for human-computer interaction. In educational settings, FER can help teachers understand the emotions of their students in a classroom, allowing them to use this feedback to enhance the learning experience [2].

In recent years, FER has been dominated by deep learning systems, primarily leveraging deep Convolutional Networks such as ResNets [3] and AlexNets [4]. Vision Transformers [5], widely used in Natural Language Processing (NLP), have also been applied to FER, significantly enhancing performance. In [6], MobileFaceNet [7] was employed as a backbone for feature extraction from images, and a Dual Direction Attention Network (DDAN) was proposed. DDAN generates attention maps from both vertical and horizontal orientations, guiding the model to focus on specific facial features and providing detailed feature representation. An attention loss mechanism was introduced to ensure that different attention heads concentrate on distinct features.

In [8], the Local Multi-Headed Channel (LHC) module was proposed, which can be integrated into existing CNN architectures to provide channel-wise self-attention. Pecoraro et.al incorporated LHC into ResNet34, introducing LHC-Net for facial emotion classification. A similar CNN-based approach with an attention mechanism is seen in [9], where ResNet was used as a backbone along with spatial and channel attention mechanisms and two loss functions. In [10], a two-stream feature extraction method using an image backbone and a facial landmark detector was studied, where image and landmark features are divided into non-overlapping windows for cross-attention and then fed to a transformer block for classification. This method is computationally more cost-effective compared to [11]. In [12], Vision Transformers (ViTs) with Multi-View Complementary Prompters (MCPs) were proposed for FER. ViT extracts facial features while MCP enhances performance by combining landmark features. A novel ensemble approach is introduced in [13], advocating the use of the ResMaskingNetwork in conjunction with other networks. In [14], a multi-task learning (MTL) approach integrates the EmoAffectNet [15] and EffNet-B2 models, introducing a coupling loss technique to mitigate negative transfer effects and facilitate improved learning of facial features across multiple datasets.

In this paper, we propose ResEmoteNet, a novel neural network architecture for facial emotion recognition. ResEmoteNet integrates Convolutional Neural Networks, Residual connections, and the Squeeze and Excitation network [16] to effectively capture facial emotions. We evaluated the proposed network using three open-source databases: FER2013 [17], RAF-DB [18], and AffectNet [19]. ResEmoteNet demonstrated state-of-the-art performance across all three databases.

The remaining of this paper is organized as follows: Section II presents the architecture of the proposed ResEmoteNet. Section III introduces the benchmark datasets and presents

Arnab Kumar Roy is with Sikkim Manipal Institute of Technology (SMIT), Sikkim, India - 737136 (e-mail: arnab_202000152@smit.smu.edu.in), Hemant Kumar Kathania, Adhitiya Sharma, and Md. Sarfaraj Alam Ansari are with National Institute of Technology Sikkim, India - 737139 (e-mail: hemant.ece@nitsikkim.ac.in, b180078@nitsikkim.ac.in, sarfaraj@nitsikkim.ac.in), Abhishek Dey is with Bay Area Advanced Analytics India (P) Ltd, A Kaliber.AI company, Guwahati, India - 781039 (e-mail: abhishek@kaliberlabs.com)

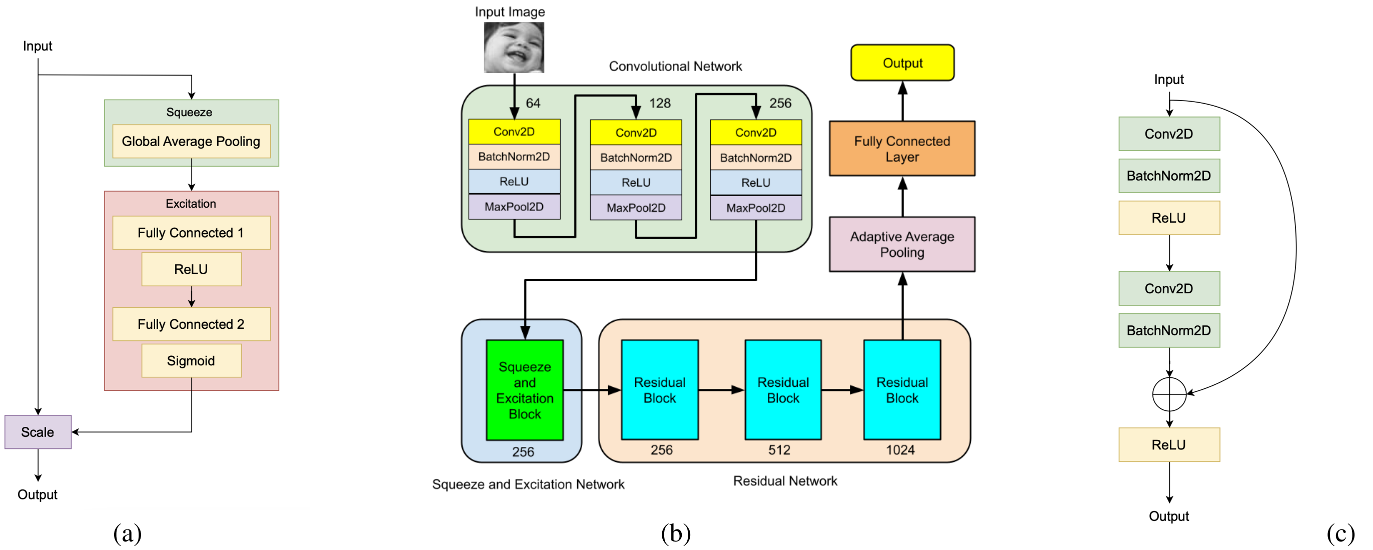


Fig. 1: (a) Architecture of Squeeze and Excitation Block, (b) Overall Architecture of Proposed ResEmoteNet (c) Architecture of Residual Block

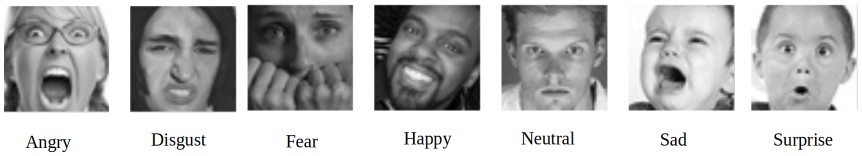


Fig. 2: Representative images of the 7 emotion classes: Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise, showcasing the varied facial expressions in the datasets.

the implementation details of the experiments. Section IV discusses the experimental results and compares with state-of-the-art methods and finally the paper is concluded in Section V.

II. Proposed Method

This section presents our proposed **ResEmoteNet** framework, with a comprehensive architecture illustrated in Fig. 1 (b). The framework integrates a simple Convolutional Neural Network (CNN) block, complemented by the Squeeze and Excitation (SE) block and reinforced by multiple Residual blocks, forming a robust and efficient network.

A. Convolutional Network

Our architecture features a Convolutional Neural Network (CNN) module, comprising three convolutional layers that progressively extract hierarchical features from the input data. To optimize learning and training efficiency, each layer is followed by batch normalization, which standardizes inputs and stabilizes the learning process. Furthermore, max-pooling is applied to reduce spatial dimensions, thereby reducing computational overhead and introducing translational invariance to boost robustness. These layers serve as the basis of our feature extraction process. Mathematically, the Convolutional Network's (CNet) feature extraction process is expressed as follows:

$$X_{FE} = CNet(X) \tag{1}$$

where X is the raw image sample and X_{FE} is the output of the feature map from the CNet.

B. Squeeze and Excitation Network

Squeeze and Excitation Network (SENet) is incorporated into our methodology to boost the representational power of convolutional neural networks. At the heart of SENet lies the SE block, a key component that models the relationships between convolutional channels. It performs two primary functions: Squeeze, which uses global average pooling to condense spatial data from each channel into a global descriptor, and Excitation, which employs a sigmoid-activated gating mechanism to capture channel dependencies. SENet's approach allows the network to learn a series of attention weights, highlighting the importance of each input element for the network's output. The architecture of the Squeeze and Excitation block is shown in Fig. 1 (a).

Let X_{FE} be the input for the squeeze operation which can be expressed as:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{FE}(c, i, j) \tag{2}$$

Here, H and W are the height and widths of the feature maps, respectively, and $X_{FE}(c, i, j)$ denotes the activation at position (i, j) in channel c.

The squeezed output z is then processed through two fully connected layers: a dimensionality-reduction layer followed by a dimensionality-expansion layer, with a Rectified Linear Unit (ReLU) activation in between. The excitation operation can be formulated as:

$$X_{S1} = ReLU(W_1 z) \tag{3}$$

Here, W_1 is the weight matrix that reduces the dimensionality. X_{S1} is the output from first dimension-reduction layer.

$$X_{S2} = W_2 X_{S1} \tag{4}$$

Where, W_2 is the weight matrix that expands it back to the original number of channels i.e., $c.X_{S2}$ is the output from first dimension-expansion layer.

$$s = sigmoid(X_{S2}) \tag{5}$$

The per-channel modulation weights, derived from the excitation phase output s, are employed to adjust the original input feature maps X_{FE}. This scaling operation is done on each element individually and is represented by Y:

$$Y = s \cdot X_{FE} \tag{6}$$

C. Residual Network

Residual Networks (ResNets) are a significant innovation in deep learning, particularly in fields that involve training extremely deep neural networks. He et al. [3] introduced ResNets, which efficiently tackle the common issues of vanishing and exploding gradients in neural networks. ResNets' main innovation is the addition of the residual block, which includes a shortcut connection to skip one or more layers. Mathematically, the function in a residual block can be defined as:

$$F(x) = H(x) + x \tag{7}$$

In a residual block, x is the input, $H(x)$ is the output from stacked layers, and $F(x)$ is the final output. Adding x to $H(x)$ enables the network to learn identity mapping, ensuring deeper layers perform as well as shallow ones, addressing degradation. This allows training deeper networks than before, improving tasks like image classification and object detection on benchmark datasets like ImageNet [20]. Fig. 1 (c) shows the architecture of a residual block.

D. Adaptive Average Pooling

Adaptive Average Pooling (AAP) is a technique that was first introduced in 2018 [21]. AAP is a type of pooling layer used in CNNs that enables the aggregation of input information into a constant output size, regardless of the original input dimensions. AAP adjusts kernel size and stride to reach a specific output size, instead of reducing spatial dimensions like traditional pooling methods. It ensures consistent output dimensions in various datasets and layers.

Considering X_{RB} as the output feature map from the residual block and $\tilde{X}_{FE}$ be the output from the AAP operation, this operation can be expressed as:

$$\tilde{X}_{FE} = AAP(X_{RB}) \tag{8}$$

$\tilde{X}_{FE}$ is finally fed to the classifier that outputs a probability distribution over the possible facial expressions.

$$P = Classifier(\tilde{X}_{FE}) \tag{9}$$

$P \in \mathbb{R}^N$, where N is the number of facial emotion classes. $Classifier$ is the Linear layer that helps to classify image based on output of the network.

III. Dataset and Experimental details

A. Dataset

In this sub-section, we provide a brief overview of the datasets used in this study. We conducted our experimental studies on three popular Facial Emotion Recognition datasets namely **FER2013 [17], RAF-DB [18]** and **AffectNet [19]**. A comparison of the three datasets is presented in Table I, detailing their number of channels, image sizes, number of samples and number of classes. Our facial emotion recognition task involves identifying seven fundamental emotions: Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise. Visual examples of each emotion are displayed in Fig. 2. To ensure a comprehensive analysis, we have also included a class-wise breakdown of the train-test distribution for each dataset in Table II.

TABLE I: Dataset comparison

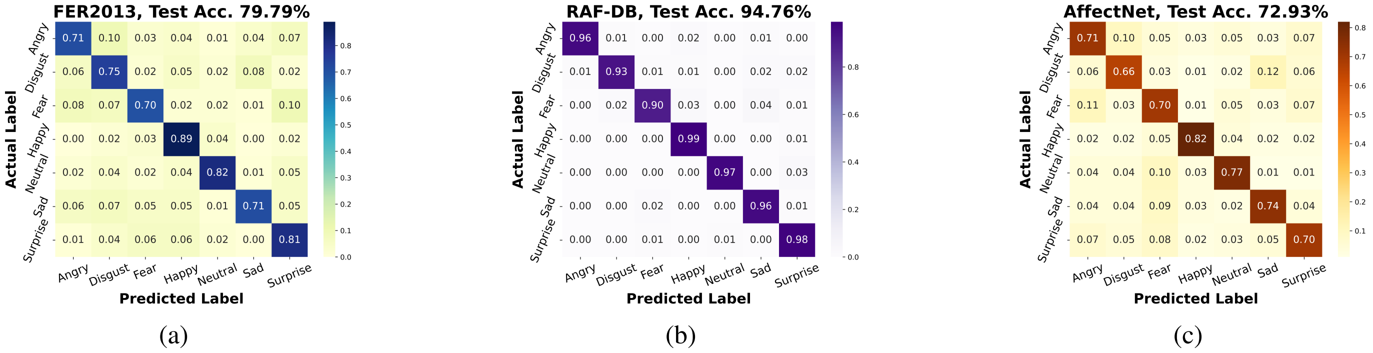
Characteristics	**FER2013**	**RAF-DB**	**AffectNet**
Number of channels	1	3	3
Image size	48 × 48	100 × 100	224 × 224
Total Samples	35,887	15,339	287,401
Number of classes	7	7	7

TABLE II: Class wise data distribution across three datasets: FER2013, RAF-DB and AffectNet.

Class Names	**FER2013**		**RAF-DB**		**AffectNet**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
Angry	3995	491	705	162	24882	500
Disgust	436	416	717	160	3803	500
Fear	4097	626	281	74	6378	500
Happy	7215	594	4772	1185	134415	500
Neutral	4965	528	2524	680	74874	500
Sad	4830	879	1982	478	25459	500
Surprise	3171	55	1290	329	14090	500
Total Samples	**28709**	**3589**	**12271**	**3068**	**283901**	**3500**

B. Experimental Details

In this study, we employed a consistent experimental setup across three facial emotion datasets, namely FER2013, RAF-DB, and AffectNet. Our model architecture, ResEmoteNet, was trained separately on each dataset using a fixed set of hyper parameters: a batch size of 16, 80 training epochs, Cross-Entropy Loss [22] as the cost function, and Stochastic Gradient Descent (SGD) [23] as the optimizer. To enhance the robustness of our model, we applied Random Horizontal Flip as a data augmentation technique. The initial learning rate was set to 1×10^{-3}, with a learning rate scheduler that reduced the rate by a factor of 0.1 upon reaching a plateau, defined as no improvement in model accuracy for a predetermined number



(a) (b) (c)

Fig. 3: Confusion matrices of ResEmoteNet across three databases: (a) for FER2013, (b) for RAF-DB and (c) for AffectNet.

of epochs. Our experiments were executed on two distinct hardware configurations: a MacBook Pro (M2 Pro-Chip with 10-core CPU and 16-core GPU) and a NVIDIA Tesla P100 GPU provided by Kaggle. The implementation was done using PyTorch [24]. For our facial emotion recognition tasks, we employ Accuracy (%) as the evaluation metric, defined as:

$$\text{Accuracy } (\%) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100 \qquad (10)$$

where TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative.

IV. Results and Discussion

In this section, we present the experimental results conducted on three widely used benchmark datasets, FER2013, RAF-DB, and AffectNet. We evaluated our proposed method, ResEmoteNet, on these datasets and compared it with other state-of-the-art methods, as shown in Table III. The results demonstrate that our proposed method outperforms the current state-of-the-art techniques. We present the confusion matrices of the ResEmoteNet for each of the datasets depicting their class wise confusions on the respective test sets in Fig. 3,

TABLE III: Test Accuracy (%) comparison of ResEmoteNet with existing state-of-the-art methods across three datasets: FER2013, RAF-DB and AffectNet.

	Accuracy in %		
Method	**FER2013**	**RAF-DB**	**AffectNet**
LHC-Net [8]	74.42	-	-
Local Learning Deep+BOW [25]	75.42	-	-
Segmentation VGG-19 [26]	75.97	-	-
EmoNeXt [27]	76.12	-	-
Ensemble ResMaskingNet [13]	76.82	-	-
POSTER++ [10]	-	92.21	67.49
DCJT [28]	-	92.24	-
DDAMFN++ [6]	-	92.34	67.36
ARBEx [29]	-	92.47	-
S2D [12]	-	92.57	67.62
C MT EffNet-B2 [14]	-	-	68.9
C MT EmoAffectNet [14]	-	-	69.4
Proposed ResEmoteNet	**79.79**	**94.76**	**72.93**

A. FER2013

Working with FER2013 has always been difficult due to inaccurate labeling, absence of faces in some images, and its data distribution. This factor has contributed to poor performance in this dataset. However, in contrast to previous studies, ResEmoteNet achieved a classification accuracy of 79.79%, representing a 2.97% absolute improvement over Ensemble ResMaskingNet [13].

B. RAF-DB

RAF-DB was selected due to its variety of real-life scenarios and difficult challenges. It is a significant dataset to assess FER techniques across different variables like pose, lighting, and occlusion. Our model ResEmoteNet achieved a classification accuracy of 94.76% having a 2.19% absolute improvement over S2D [12].

C. AffectNet

AffectNet with 7 emotions was selected for being one of the biggest facial expression recognition datasets that is publicly accessible along with a broad assortment of annotated facial pictures includes a variety of emotions. Our model ResEmoteNet achieved a classification accuracy of 72.93% having a 3.53% absolute improvement over EmoAffectNet [14].

V. Conclusion

In this paper, we introduced ResEmoteNet, a novel neural network architecture designed to address the challenging task of facial emotion recognition. Our model integrates a combination of three distinct networks: Convolutional Neural Network, Squeeze and Excitation network and residual network. The integration of these networks allows ResEmoteNet to effectively capture and interpret complex emotional expressions from facial images. To evaluate the performance of our proposed ResEmoteNet, we conducted extensive experiments using three widely recognized benchmark datasets: FER2013, RAF-DB, and AffectNet. Our experimental results demonstrate that ResEmoteNet consistently outperforms existing state-of-the-art models across all these datasets. These results highlight the effectiveness of our approach in accurately recognizing facial emotions, offering significant advancements in the field of facial emotion recognition. Future work will explore further enhancements to ResEmoteNet and its applications in real-world scenarios.

VI. Acknowledgement

This work was supported by the iHub DivyaSampark IIT Roorkee.

References

[1] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma, "Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders," *Journal of neuroscience methods*, vol. 200, no. 2, pp. 237–256, 2011.

[2] N. Bosch, S. K. D'Mello, R. S. Baker, J. Ocumpaugh, V. Shute, M. Ventura, L. Wang, and W. Zhao, "Detecting student emotions in computer-enabled classrooms.," in *International Joint Conference on Artificial Intelligence*, vol. 16, pp. 4125–4129, 2016.

[3] 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*, pp. 770–778, 2016.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *International Conference on Learning Representations ICLR*, 2021.

[6] S. Zhang, Y. Zhang, Y. Zhang, Y. Wang, and Z. Song, "A dual-direction attention mixed feature network for facial expression recognition," *MDPI Electronics*, vol. 12, no. 17, p. 3595, 2023.

[7] S. Chen, Y. Liu, X. Gao, and Z. Han, "Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices," in *Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13*, pp. 428–438, Springer, 2018.

[8] R. Pecoraro, V. Basile, and V. Bono, "Local multi-head channel self-attention for facial expression recognition," *MDPI Information*, vol. 13, no. 9, p. 419, 2022.

[9] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: Multi-head cross attention network for facial expression recognition," *MDPI Biomimetics*, vol. 8, no. 2, p. 199, 2023.

[10] J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, and A. Huang, "Poster++: A simpler and stronger facial expression recognition network," *arXiv preprint arXiv:2301.12149*, 2023.

[11] C. Zheng, M. Mendieta, and C. Chen, "Poster: A pyramid cross-fusion transformer network for facial expression recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3146–3155, 2023.

[12] Y. Chen, J. Li, S. Shan, M. Wang, and R. Hong, "From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos," *arXiv preprint arXiv:2312.05447*, 2023.

[13] L. Pham, T. H. Vu, and T. A. Tran, "Facial expression recognition using residual masking network," in *2020 25Th international conference on pattern recognition (ICPR)*, pp. 4513–4519, IEEE, 2021.

[14] D. Kollias, V. Sharmanska, and S. Zafeiriou, "Distribution matching for multi-task learning of classification tasks: a large-scale study on faces & beyond," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 2813–2821, 2024.

[15] E. Ryumina, D. Dresvyanskiy, and A. Karpov, "In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study," *Neurocomputing Elsevier*, 2022.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[17] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, pp. 117–124, Springer, 2013.

[18] 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*, pp. 2852–2861, 2017.

[19] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[21] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8759–8768, 2018.

[22] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.

[23] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[24] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[25] M.-I. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64827–64836, 2019.

[26] S. Vignesh, M. Savithadevi, M. Sridevi, and R. Sridhar, "A novel facial emotion recognition model using segmentation vgg-19 architecture," *International Journal of Information Technology*, vol. 15, no. 4, pp. 1777–1787, 2023.

[27] Y. El Boudouri and A. Bohi, "Emonext: an adapted convnext for facial emotion recognition," in *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, IEEE, 2023.

[28] C. Yu, D. Zhang, W. Zou, and M. Li, "Joint training on multiple datasets with inconsistent labeling criteria for facial expression recognition," *IEEE Transactions on Affective Computing*, 2024.

[29] A. T. Wasi, K. Šerbetar, R. Islam, T. H. Rafi, and D.-K. Chae, "Arbex: Attentive feature extraction with reliability balancing for robust facial expression learning," *arXiv preprint arXiv:2305.01486*, 2023.