

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

arXiv:2409.10545v2 [cs.CV] 23 Nov 2024

Abstract—The human face is a silent communicator, expressing emotions and thoughts through its 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 four open-source databases: FER2013, RAF-DB, AffectNet-7 and ExpW, achieving accuracies of 79.79%, 94.76%, 72.39% and 75.67% respectively. The proposed network outperforms state-of-the-art models across all four 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].

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)

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 introduced to add channel-wise self-attention to existing CNN architectures. Pecoraro et al. incorporated LHC into ResNet34, creating LHC-Net for facial emotion classification. A similar approach is found in [9], where ResNet is combined with spatial and channel attention mechanisms and two loss functions. In [10], a two-stream feature extraction method using an image backbone and facial landmark detector was proposed, utilizing cross-attention for classification. This method is more computationally efficient than [11].

In [12], Vision Transformers (ViTs) with Multi-View Complementary Prompts (MCPs) were proposed for FER, where ViT extracts facial features and MCP enhances performance by combining landmark features. An ensemble approach is introduced in [13], using the ResMaskingNetwork with other networks. In [14], a multi-task learning approach integrates the EmoAffectNet and EffNet-B2 models, using a coupling loss to improve facial feature learning across datasets. In [15], an attention mechanism with self-supervised learning, diffusion-based denoising, and restoration improves performance in noisy environments like under-display cameras (UDC). Group-GAN, proposed in [16], includes an extractor, generator, and two discriminators to handle weak emotions by converting them to peak emotions, improving performance. Finally, [17] introduces HTNet for micro-expression recognition, leveraging local temporal, local, and global semantic features, demonstrating enhanced results across four publicly available 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 [18] to effectively capture facial emotions. We evaluated the proposed network using four open-source databases: FER2013 [19], RAF-DB [20], AffectNet-7 [21] and ExpW (Expressions in

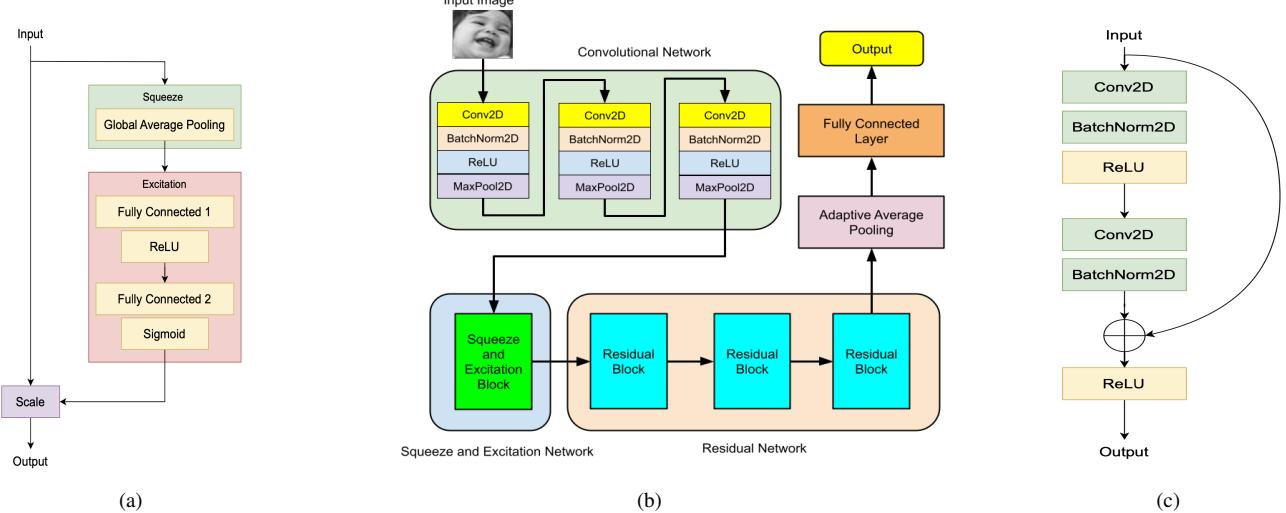


Fig. 1: (a) Architecture of Squeeze and Excitation Block, (b) Overall Architecture of Proposed ResEmoteNet (c) Architecture of Residual Block

the Wild) [22]. ResEmoteNet demonstrated state-of-the-art performance across all four databases.

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 includes a CNN module with three convolutional layers for hierarchical feature extraction, followed by batch normalization to stabilize learning and enhance training efficiency. Max-pooling is applied after each layer to reduce spatial dimensions, lowering computational costs and introducing translational invariance for improved robustness. These layers form the foundation of our feature extraction process. Mathematically, the Convolutional Network's (CNet) feature extraction process is expressed as follows:

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

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:

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

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:

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

$$X_{S2} = W_2 X_{S1} \quad (4)$$

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

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

$$Y = s \cdot X_{FE} \quad (6)$$

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:

$$F(x) = H(x) + x \quad (7)$$

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

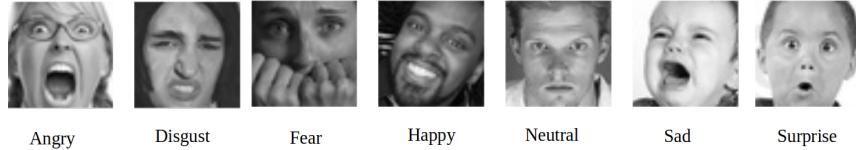


Fig. 2: Representative images of the 7 emotion classes: Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise, showcasing the varied facial expressions in the datasets.

$H(x)$ 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 [23]. 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 [24]. 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:

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

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

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

$P \in \mathbb{R}^N$, 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 four popular Facial Emotion Recognition datasets namely **FER2013** [19], **RAF-DB** [20], **AffectNet-7** [21] and **ExpW (Expressions in the Wild)** [22]. A comparison of the four 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-7	ExpW
No. of channels	1	3	3	3
Image size	48×48	100×100	224×224	224×224
Total Samples	35,887	15,339	287,401	91,793
No. of classes	7	7	7	7

TABLE II: Class wise data distribution across four datasets: FER2013, RAF-DB, AffectNet-7 and ExpW

Class	FER2013		RAF-DB		AffectNet-7		ExpW	
	Train	Test	Train	Test	Train	Test	Train	Test
Angry	3995	491	705	162	24882	500	2569	364
Disgust	436	416	717	160	3803	500	2796	396
Fear	4097	626	281	74	6378	500	761	108
Happy	7215	594	4772	1185	134415	500	21375	3024
Neutral	4965	528	2524	680	74874	500	24418	3454
Sad	4830	879	1982	478	25459	500	7391	1046
Surprise	3171	55	1290	329	14090	500	4943	699
Total	28709	3589	12271	3068	283901	3500	64253	9091

B. Experimental Details

In this study, we employed a consistent experimental setup across four facial emotion datasets: FER2013, RAF-DB, AffectNet and ExpW. ResEmoteNet was trained on each dataset using carefully selected hyperparameters, based on sensitivity analysis experiments. We evaluated batch sizes of 16, 32, 64, and 128, and found that a batch size of 16 consistently yielded superior performance, likely because more frequent weight updates facilitated faster convergence during training. Similarly, we experimented with epochs ranging from 40 to 80, with 80 epochs being the optimal choice, as the model's performance saturated around this point, and further increases yielded minimal gains in accuracy. Thus, the final configuration employed a batch size of 16 and 80 training epochs across all datasets, using Cross-Entropy Loss [25] as the cost function and Stochastic Gradient Descent (SGD) [26] as the optimizer. As the classes in all the datasets are not balanced as seen from Table II, we used a class-weight biased method for training, where the classes with fewer samples were given more weightage in the loss function. We applied Random Horizontal Flip for data augmentation and used a learning rate of 1×10^{-3} , with a scheduler that reduced the rate by 0.1 when performance plateaued. 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 [27]. For our facial emotion recognition tasks, we employ Accuracy (%) as the evaluation metric, defined as:

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

where TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative. ResEmoteNet was trained for 3 hours for FER2013, 2 hours for RAF-DB, 6 hours for AffectNet-7 and 4.5 hours for ExpW, with inference times of 1.4ms, 3.2ms, 3.6ms and 1.9ms per image respectively.

IV. RESULTS AND DISCUSSION

In this section, we present the experimental results conducted on four widely used benchmark datasets, FER2013,

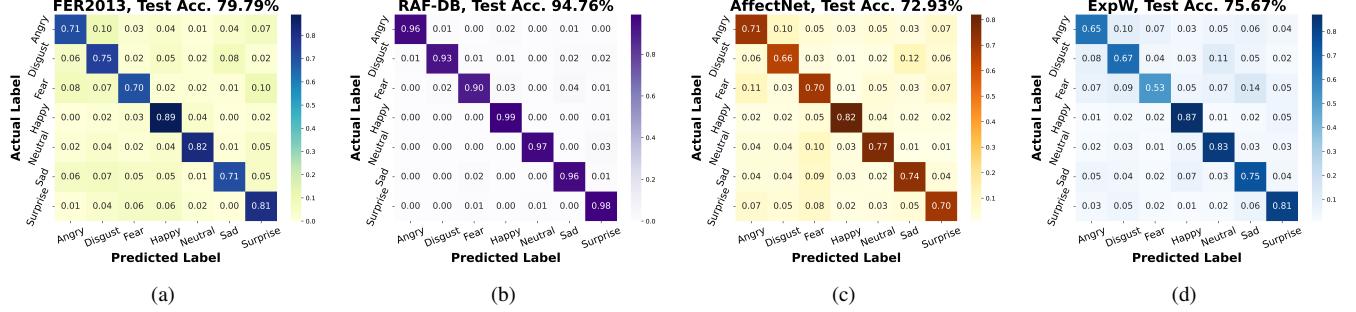


Fig. 3: Confusion matrices of ResEmoteNet across four databases: (a) for FER2013, (b) for RAF-DB (c) for AffectNet-7 and (d) for ExpW.

RAF-DB, AffectNet-7 and ExpW. We evaluated our proposed method, ResEmoteNet, on these datasets and tabulated in Table III. We also 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.

A. Model Performance Across Datasets: Comparison with Prior Studies

In Table III, we compared the performance of the proposed method with other state-of-the-art methods, the results demonstrate that our proposed method outperforms the current state-of-the-art techniques. Working with FER2013 is challenging due to inaccurate labeling, absence of faces in some images, and imbalanced data distribution, yet ResEmoteNet achieved a classification accuracy of 79.79%, a 2.97% absolute improvement over Ensemble ResMaskingNet [13]. On RAF-DB, which presents real-life challenges such as pose, lighting, and occlusion, ResEmoteNet attained a classification accuracy of 94.76%, improving by 2.19% over S2D [12]. On AffectNet-7, a large-scale dataset with 7 emotions and diverse annotations, our model achieved a classification accuracy of 72.93%, showing a 3.53% improvement over EmoAffectNet [14]. Lastly, on ExpW, which presents a broad range of facial expressions in uncontrolled, real world settings, ResEmoteNet achieved a classification accuracy of 75.67%, showing a 2.19% improvement over APViT [28].

TABLE III: Test Accuracy (%) comparison of ResEmoteNet with existing state-of-the-art methods across four datasets: FER2013, RAF-DB, AffectNet-7 and ExpW.

Method	Accuracy in %			
	FER2013	RAF-DB	AffectNet-7	ExpW
Seg. VGG-19 [29]	75.97	-	-	-
EmoNeXt [30]	76.12	-	-	-
En. ResMaskingNet [13]	76.82	-	-	-
SEResNet [31]	-	83.37	56.54	-
Arm [32]	-	90.42	62.5	-
APViT [28]	-	91.98	66.91	73.48
ARBEx [33]	-	92.47	-	-
S2D [12]	-	92.57	67.62	-
C MT EmoAffectNet [14]	-	-	69.4	-
AGLRLS [34]	-	-	-	73
SchiNet [35]	-	-	-	73.10
Proposed ResEmoteNet	79.79	94.76	72.93	75.67

B. Ablation Study

We conducted an ablation study on the RAF-DB dataset to evaluate the contribution of key components: Convolutional layers, SE blocks, and Residual Blocks within the ResEmoteNet architecture. The results in Table IV demonstrate

the impact on accuracy and loss when these components are removed.

Removing the SE block resulted in a minor accuracy drop from 94.76% to 94.18%, while eliminating Residual Block 1 (with 256 channels) led to a substantial reduction in accuracy, from 94.76% to 75.39%, highlighting the critical role of residual connections. Similarly, removing convolutional layers, such as CNN 64 and CNN 64, 128, caused accuracy decreases of 11.25% and 6.49%, respectively.

TABLE IV: Component-wise study of ResEmoteNet in RAF-DB Dataset. Note that Config C1 denotes our best ResEmoteNet; CNN stands for Convolutional Neural Network, SE stands for Squeeze-and-Excitation Block, RB stands for Residual Block respectively.

Config	Components			Removed Component(s)	Accuracy (%)	Loss (%)
	CNN	SE	RB			
C1	64, 128, 256	256	256, 512, 1024	-	94.76	3.73
C2	64, 128, 256	-	256, 512, 1024	SE (256)	94.18 (-0.58)	16.83 (+13.1)
C3	128, 256	256	256, 512, 1024	CNN (64)	83.51 (-11.25)	7.19 (+3.46)
C4	256	256	256, 512, 1024	CNN (64, 128)	88.27 (-6.49)	5.32 (+1.59)
C5	64, 128, 256	256	512, 1024	RB (256)	75.39 (-19.37)	10.2 (+6.47)
C6	64, 128, 256	256	1024	RB (256, 512)	82.77 (-11.99)	8.8 (+5.07)

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 four widely recognized benchmark datasets: FER2013, RAF-DB, AffectNet-7 and ExpW. Our experimental results demonstrate that ResEmoteNet consistently outperforms existing state-of-the-art models across all these datasets. Additionally, an ablation study on the RAF-DB dataset confirmed the critical role of the residual and convolutional layers in maintaining the model's performance. 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] Z. Wang, K. Zhang, and R. Sankaranarayana, "Lrdif: Diffusion models for under-display camera emotion recognition," *arXiv preprint arXiv:2402.00250*, 2024.
- [16] W. Niu, K. Zhang, D. Li, and W. Luo, "Four-player groupgan for weak expression recognition via latent expression magnification," *Knowledge-Based Systems*, vol. 251, p. 109304, 2022.
- [17] Z. Wang, K. Zhang, W. Luo, and R. Sankaranarayana, "Htnet for micro-expression recognition," *Neurocomputing*, vol. 602, p. 128196, 2024.
- [18] 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.
- [19] 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.
- [20] 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.
- [21] 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.
- [22] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *International Journal of Computer Vision*, vol. 126, pp. 550–569, 2018.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.
- [27] 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.
- [28] F. Xue, Q. Wang, Z. Tan, Z. Ma, and G. Guo, "Vision transformer with attentive pooling for robust facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 3244–3256, 2022.
- [29] 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.
- [30] 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.
- [31] Z.-Y. Huang, C.-C. Chiang, J.-H. Chen, Y.-C. Chen, H.-L. Chung, Y.-P. Cai, and H.-C. Hsu, "A study on computer vision for facial emotion recognition," *Scientific Reports*, vol. 13, no. 1, p. 8425, 2023.
- [32] J. Shi, S. Zhu, and Z. Liang, "Learning to amend facial expression representation via de-albino and affinity," *arXiv preprint arXiv:2103.10189*, 2021.
- [33] 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.
- [34] Y. Gao, Y. Xie, Z. Z. Hu, T. Chen, and L. Lin, "Adaptive global-local representation learning and selection for cross-domain facial expression recognition," *IEEE Transactions on Multimedia*, 2024.
- [35] M. Bishay, P. Palasek, S. Priebe, and I. Patras, "Schinet: Automatic estimation of symptoms of schizophrenia from facial behaviour analysis," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 949–961, 2019.