2024 IEEE International Conference on Image Processing (ICIP) | 979-8-3503-4939-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICIP51287.2024.10648251

A SELF-SUPERVISED DIFFUSION FRAMEWORK FOR FACIAL EMOTION RECOGNITION

Saif Hassan[1,2], *Mohib Ullah*[1], *Ali Shariq Imran*[1], *Ghulam Mujtaba*[2], *Muhammad Mudassar Yamin*[3], *Ehtesham Hashmi*[3], *Faouzi Alaya Cheikh*[1], *Azeddine Beghdadi*[1,4]

[1] Intelligent Systems and Analytics (ISA) Research Group, Department of Computer Science (IDI), Norwegian University of Science & Technology (NTNU), Gjøvik, 2815, Norway.
[2] Department of Computer Science, Sukkur IBA University, Sukkur, 65200, Pakistan
[3] Department of Information Security and Communication Technology, NTNU Norway.
[4] Université Sorbonne Paris Nord, Paris France.

ABSTRACT

In this paper, we introduced a novel Facial Emotion Recognition (FER) framework that utilizes a diffusion-based approach and an attention mechanism. The model is efficiently trained through self-supervised learning, leveraging labeled and unlabelled data. The proposed framework has been rigorously tested on the FER2013 and AffectNet datasets, achieving promising accuracies of 67.2% and 68.1%, respectively. The quantitative results not only surpass the performance of existing state-of-the-art FER models but also demonstrate the synergistic effect of combining diffusion-based modeling with self-supervised learning and attention mechanisms within a solid architectural framework. Our approach sets a new benchmark in the field, offering a significant step forward in the accurate and efficient recognition of facial expressions.

Index Terms— Diffusion model, self-supervised learning, attention mechanism, facial emotion recognition.

1. INTRODUCTION

In recent years, a surge in refining Facial Emotion Recognition (FER) systems has been driven by the dual forces of deep learning and computer vision [1, 2, 3, 4]. Researchers have proposed various FER algorithms, addressing the problem through diverse approaches. For example, Mao et al. [5] combined focal loss and CosFace loss with ResNet-18 model. Similarly, the work of Luo et al. [6] introduced a data augmentation method, that marked a significant milestone in enhancing model performance. The study of Mao et al. [7] introduces a model named POSTER++. POSTER++ successfully achieved 63.77% accuracy on AffectNet dataset. Similarly, Zhang et al. [8] presented a Dual-Direction Attention Mixed Feature Network (DDAMFN) that utilizes dual-direction attention mechanisms to enhance the model's ability to detect expressions accurately. The transition from static images to dynamic video for FER is addressed by Chen et al. [9]. They explored the adaptation of image-based FER models to video inputs, emphasizing the importance of temporal dynamics and facial landmarks in accurately recognizing expressions over time. Compared to visual emotion recognition, Khan et al. [10] introduce a lightweight deep learning model to recognize Speech emotions (SEs). They employed an adaptive wavelet thresholding for pre-processing and using spectrograms to enhance feature extraction. Furthermore, Savchenko et al. [11] used FER in the context of online education. They presented a neural network designed to classify a range of emotions and engagement levels based solely on facial expressions captured during learning sessions. The advent of diffusion models introduces a paradigm shift in image processing, treating images as dynamic systems that evolve over time. These models, particularly denoising diffusion-based conditional generative models [12], offer a novel perspective on generating high-quality, denoised samples. This paper capitalizes on the diffusion model framework, especially the recent advancements exemplified by the CARD model [13] combined with the Convolutional Block Attention Module (CBAM). In a nutshell, the contribution of our work is three folds:

- We leveraged the CARD model [13], employing a diffusion-based strategy, to achieve precise facial expression recognition.
- We integrate a spatial and channel attention module [14, 15] in our framework, which aids in extracting relevant and discriminative features.
- We adopted a self-supervised learning approach to train our framework end-to-end using both the label and unlabelled data.

Our framework is modular in terms of backbone architecture, supporting any CNN [16] or Vision Transformer [17] based backbone. Our experiments reveal notable improvements in performance, surpassing previous state-of-the-art methods. The rest of the paper is organized in the following sections.

Details of our proposed framework including the training strategy and attention mechansims are explained in section 2. Description of the dataset, evaluation metrics are discussed in section 3. The implementation details including data preprocessing, and model training are elaborated in section 4. The quantitative results and the ablation study are discussed in section 5. Section 6 concluded the paper with the final remarks and future directions.

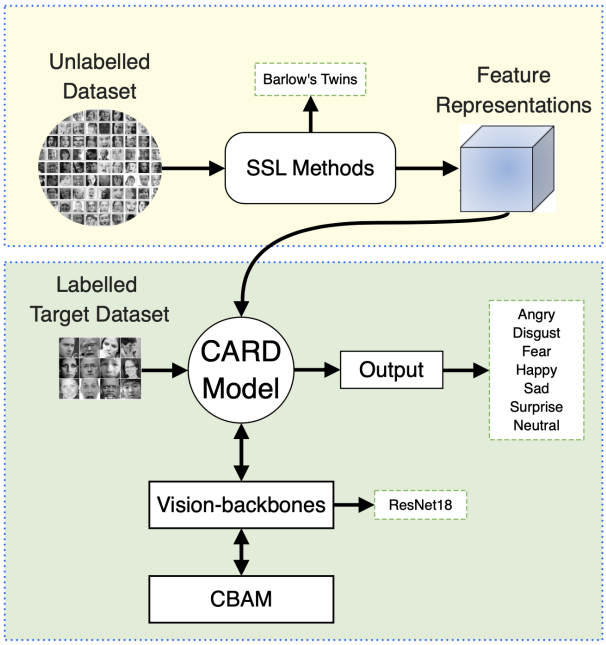

Fig. 1. The proposed framework leverages labeled and labeled data for training the model. The backbone is modular. and the features are refined through spatial and channel attention.

2. METHODOLOGY

The block diagram of our proposed framework is shown in figure 1. At the core of our methodology lies the CARD model [13], a novel diffusion-based technique that significantly improves the precision of facial expression recognition. We employed the Barlow Twins self-supervised learning (SSL) method [18], which facilitates a deeper understanding of the intricate patterns in facial expressions. The Convolutional Block Attention Module (CBAM) is integrated to ensure the model focuses on the most relevant and discriminative features within the input images. In the following subsections, each component of our framework is explained in detail.

2.1. CARD Model

We utilized the CARD model [13] to define the conditional distribution $p(y|x, D)$. This choice was based on the CARD model's capacity to model implicit distributions through stochastic processes, particularly leveraging the capabilities inherent in diffusion models. The process can be summarized in two key points: **Convergence to Stationary Distribution** With a sufficiently large number of timesteps T, the noising process $q(x_T|x_0)$ converges to a stationary distribution $p(x_T)$ regardless of the initial distribution $q(x_0)$. Mathematically:

$$\lim_{T\to\infty} q(x_T|x_0) = p(x_T) \tag{1}$$

This indicates the ability to transition from any initial distribution $q(x_0)$ to a stationary distribution $p(x_T)$ through the noising process. **Forward and Reverse Process Functional Form** With large T and a small enough noise schedule $\{\beta_t\}_{t=1}^{T}$, the forward and reverse diffusion processes share the same functional form. The product in $q(x_{t-1}|x_t) \propto q(x_t|x_{t-1})q(x_{t-1})$ is dominated by $q(x_t|x_{t-1})$, allowing the reverse process to mimic the forward process closely. The goal is to learn a function $p_\theta(x_{t-1}|x_t)$ that approximates $q(x_{t-1}|x_t)$ well, enabling the model to reverse the diffusion process from $p(x_T)$ back to any $q(x_0)$.

2.1.1. *Diffusion Denoising Process*

This process enables the generation of samples matching the true conditional distribution $p(y|x)$. The denoising diffusion probabilistic models (DDPM) framework, introduced by [12], is utilized, where a function approximator $\theta_{x,y,t}, f_\phi(x), t$, parameterized by a deep neural network, predicts the forward diffusion noise sampled for y_t. The CARD model is mathematically formulated to optimize an Evidence Lower Bound (ELBO) for the log-likelihood of the ground-truth response variable y_0 given its covariates x, assuming a sequence of intermediate predictions $y_{1:T}$ made by the diffusion model. The optimization objective is:

$$\log p_\theta(y_0|x) \geq E_{q(y_{1:T}|y_0,x)}\left[\log \frac{p_\theta(y_{0:T}|x)}{q(y_{1:T}|y_0,x)}\right] \tag{2}$$

This formulation incorporates both forward and reverse diffusion processes, leveraging a pre-trained conditional mean estimator to accurately estimate the conditional distribution $p(y|x, D)$. In the FER context, we used this process to generate samples from the conditional distribution of y given x.

2.2. Barlow Twins - Self Supervised Learning (SSL)

The Barlow Twins architecture [18] employs a self-supervised objective function that operates by comparing the feature representations of two distorted versions of the same input image. Figure 2 provides an overview of the Barlow Twins' learning paradigm. The process begins with an input batch of

images X, each of which is subjected to two distinct stochastic augmentation transformations $T \sim \tau$, resulting in two sets of distorted images Y^A and Y^B. These augmented images are then encoded through a shared-weight encoder f_θ to produce feature representations, which are further projected into an embedding space, resulting in Z^A and Z^B. The central

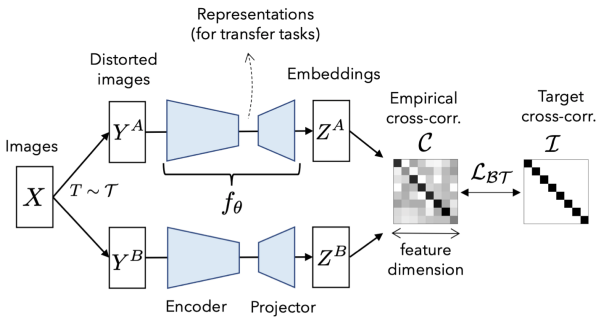

Fig. 2. Block diagram of Barlow Twins [18].

component of the Barlow Twins' objective is the empirical cross-correlation matrix C, calculated between the embedding vectors Z^A and Z^B. Each element C_{ij} of the matrix is computed as the correlation between the i^{th} feature of Z^A and the j^{th} feature of Z^B, across all samples in the batch:

$$C_{ij} = \frac{\sum_{b=1}^{B} z_{bi}^A z_{bj}^B}{\sqrt{\sum_{b=1}^{B} (z_{bi}^A)^2} \sqrt{\sum_{b=1}^{B} (z_{bj}^B)^2}} \tag{3}$$

The loss function L_{BT}, designed to make the cross-correlation matrix C close to the identity matrix I, is formulated as:

$$L_{BT} = \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2 \tag{4}$$

In equation 4, the first term prioritizes the invariance of the network by penalizing the deviation of the diagonal elements of C from 1, ensuring that the embeddings Z^A and Z^B are similar for the same features. The second term, weighted by the hyperparameter λ, enforces the redundancy reduction by penalizing the correlation between different features, thus encouraging the network to produce a diverse set of features.

2.3. Convolutional Block Attention Module (CBAM)

The Convolutional Block Attention Module (CBAM) [19] is incorporated in the backbone to boost the efficiency by selectively emphasizing or suppressing features across channels and spatial dimensions. We implemented CBAM by utilizing both channel and spatial attentions in the following manner:

Input

Consider an intermediate feature map $F \in R^{C \times H \times W}$, where C, H, and W represent the number of channels, height, and width, respectively.

Channel Attention Module (CAM)

- Global Average Pooling (GAP) and Global Max Pooling (GMP) applied on F yield $F_{avg}^C, F_{max}^C \in R^C$.
- A shared Multi-Layer Perceptron (MLP) with one hidden layer processes these vectors to output channel attention maps M_C^{avg} and M_C^{max}.
- The channel attention map M_C is obtained by:

$$M_C = \sigma(MLP(F_{avg}^C) + MLP(F_{max}^C)) \tag{5}$$

where σ is the sigmoid function.

Spatial Attention Module (SAM)

- GAP and GMP across the channel axis of F produce $F_{avg}^S, F_{max}^S \in R^{H \times W}$.
- Concatenation and convolution of these maps yield the spatial attention map M_S:

$$M_S = \sigma(f^{7\times7}([F_{avg}^S; F_{max}^S])) \tag{6}$$

with $f^{7\times7}$ denoting a convolution operation with a 7×7 filter.

Output

The refined feature map F' is computed by:

$$F' = M_S \odot (M_C \odot F) \tag{7}$$

where $\odot$ signifies element-wise multiplication.

2.4. Reasoning Behind Using ResNet18 for Facial Expression Recognition:

The selection of ResNet18 [20] as the backbone for facial emotion recognition is strategically enhanced by the integration of the Convolutional Block Attention Module (CBAM), offering a highly efficient and adaptable solution for capturing nuanced emotional expressions. ResNet18's architecture, known for its residual connections that effectively mitigate the vanishing gradient problem, is ideally suited for the computationally intensive task of identifying subtle facial cues across a diverse range of expressions. When augmented with CBAM, ResNet18's ability to focus on salient features is significantly amplified, allowing for a more discerning analysis of crucial expressive details by directing attention towards important spatial and channel-wise features. This combination not only boosts the model's accuracy in recognizing facial emotions by leveraging the strengths of ResNet18's deep learning capabilities and CBAM's attention mechanisms but also complements the methodological component i.e., CARD model. In a nutshell, with the CARD Model, we introduces a diffusion-based strategy for precise expression recognition,

while SSL enables the model to leverage both labeled and unlabeled data for robust feature learning. The CBAM focuses attention on the most relevant features, improving discriminability. ResNet18 acts as the architectural backbone, ensuring efficient feature extraction.

3. EXPERIMENTS

3.1. Datasets

The FER2013 [21] and AffectNet [22] datasets are fundamental resources in the field of facial emotion recognition, providing extensive data for training and evaluating machine learning models. In the following subsection 3.1.1 and 3.1.2, we detail each dataset and provide a mathematical representation suitable for integration into research.

3.1.1. FER2013 Dataset

The FER2013 dataset [21] consists of grayscale images of facial expressions categorized into seven universal emotions: Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Neutral. It contains 35,887 images, each of size 48x48 pixels. Let $D_{FER2013} = \{(x_i, y_i)\}_{i=1}^{N}$ represent the dataset, where:

- $x_i \in R^{48\times48}$ is a grayscale image.
- $y_i \in \{0, 1, 2, 3, 4, 5, 6\}$ is the label corresponding to one of the seven emotions.
- $N = 35,887$ is the total number of images in the dataset.

3.1.2. AffectNet Dataset

AffectNet [22] is one of the largest datasets for facial emotion recognition and contains more than 1 million facial images in total and around 0.4 million images manually labeled with eight emotions (the seven basic emotions plus Contempt) and the intensity of valence and arousal. The images are in color and have been manually annotated. Let $D_{AffectNet} = \{(x_j, y_j, v_j, a_j)\}_{j=1}^{M}$ represent the dataset, where:

- $x_j \in R^{W\times H\times3}$ is a color image, with W and H denoting the width and height, respectively.
- $y_j \in \{0, 1, 2, 3, 4, 5, 6, 7\}$ is the label corresponding to one of the eight emotions.
- $v_j \in R$ represents the valence intensity, and $a_j \in R$ represents the arousal intensity for the j^{th} image.
- M is the total number of annotated images considered for training or evaluation.

3.2. Evaluation Metrics

3.2.1. Patch Accuracy vs Patch Uncertainty (PAvPU)

The PAvPU metric, standing for Patch Accuracy vs. Patch Uncertainty, is designed to evaluate the performance of a model by assessing its ability to make accurate predictions with high confidence and to express uncertainty when its predictions are incorrect. This metric is particularly useful for models that provide probabilistic outputs, offering insights into the model's confidence calibration. The PAvPU metric can be defined mathematically as follows:

$$C_i = \begin{cases} 1 & \text{if } p_i \geq \tau \\ 0 & \text{otherwise} \end{cases}$$

Where:

- $C_i = 1$ indicates a confident prediction.
- $C_i = 0$ indicates an uncertain prediction.
- τ is a predefined confidence threshold.

For each prediction i, let p_i be the predicted probability associated with the predicted class. The model's confidence, C_i, can be defined as a binary value based on a threshold τ. To classify each prediction into one of four categories based on its correctness and confidence level:

- Confident and Correct (CC)
- Confident and Incorrect (CI)
- Uncertain and Correct (UC)
- Uncertain and Incorrect (UI)

The PAvPU metric is then calculated by taking the ratio of the sum of CC and UI predictions to the total number of predictions N:

$$PAvPU = \frac{\text{Count(CC)} + \text{Count(UI)}}{N}$$

Where:

- Count(CC) is the number of confident and correct predictions.
- Count(UI) is the number of uncertain and incorrect predictions.
- N is the total number of predictions.

3.2.2. Accuracy Metric for Facial Emotion Recognition

The accuracy metric quantitatively measures the model's ability to correctly identify facial expressions across all categories. It is one of the most straightforward and commonly used metrics for evaluating the performance of classification models. Here is a detailed mathematical formulation and explanation:

Accuracy Metric Formulation

Given a dataset with N instances where each instance is a facial expression image that belongs to one of C classes (emotions), the accuracy of the facial expression recognition model can be defined as the ratio of the number of correctly predicted instances to the total number of instances in the dataset. Mathematically, accuracy (Acc) can be expressed as:

$$Acc = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{8}$$

$$Acc = \frac{\sum_{i=1}^{N} 1(\hat{y}_i = y_i)}{N} \tag{9}$$

Where:

- N is the total number of instances in the dataset.
- $\hat{y}_i$ is the predicted class for the i^{th} instance.
- y_i is the actual class for the i^{th} instance.
- $1(\hat{y}_i = y_i)$ is an indicator function that equals 1 if the predicted class matches the actual class and 0 otherwise.

4. IMPLEMENTATION DETAILS

The FER2013 dataset, recognized for its inclusion of seven universal emotions, and the AffectNet dataset, which extends to eight emotions by adding contempt, form a robust basis for training and evaluating emotion recognition models [23]. These datasets are pivotal for developing models that accurately classify a wide range of human emotions. The mathematical representation of the dataset utilization within the context of a Convolutional Neural Network (CNN) for facial emotion recognition can be formulated as follows:

4.1. Data Preprocessing

We performed two particularly relevant techniques- Image Resizing and Image Normalization are performed. These techniques are critical for preparing the datasets for efficient training of deep learning models, such as those involving ResNet18 and CBAM. Let X_{FER} and $X_{AffectNet}$ represent the sets of images from FER2013 and AffectNet datasets, respectively. The preprocessing step involves normalizing and resizing the images to fit the input dimensions of the CNN:

$$X'_{FER} = preprocess(X_{FER})$$
$$X'_{AffectNet} = preprocess(X_{AffectNet})$$

4.2. Model Training

The training process involves feeding the preprocessed images into a CNN, enhanced by ResNet18 and CBAM for feature extraction and attention-based refinement, respectively. The output layer's dimensions are determined by the number of emotions E to be recognized:

$$Y_{predicted} = CNN_{ResNet18+CBAM}(X'_{FER} \cup X'_{AffectNet}; \theta) \tag{10}$$

where $Y_{predicted} \in R^{N \times E}$ and θ represents the trainable parameters of the CNN.

Given our computational setup, featuring dual Nvidia GTX Super 2060 Ti cards each with 8GB VRAM and 16GB of system RAM, we strategically selected a subset of 50,000 to 100,000 images from the AffectNet dataset for Self-Supervised Learning (SSL) using the Barlow Twins method. This decision strikes a balance between exploiting a sufficiently large and diverse dataset to ensure robust feature learning, and the constraints of our hardware resources.

4.3. Loss Calculation

The Cross-Entropy loss function $L(Y, \hat{Y})$ is utilized to compute the difference between the predicted emotion labels $\hat{Y}$ and the true labels Y:

$$L(Y, \hat{Y}) = -\sum_{i=1}^{N}\sum_{e=1}^{E} Y_{ie} \log(\hat{Y}_{ie}) \tag{11}$$

where N is the number of samples, and E is the number of emotion categories. The integration of Barlow Twins (SSL) and CARD models into this framework aims to enhance the feature space's representativeness and predict the distribution of facial expressions accurately. The Barlow Twins method, focusing on self-supervised learning, reduces redundancy in the feature representations, thereby improving the generalization capabilities of the model. Meanwhile, the CARD model provides a probabilistic understanding of facial expressions, enabling the system to handle the inherent ambiguity and subtlety of human emotions effectively. This cohesive integration of datasets, models, and methodologies underscores the system's capacity to learn from a diverse and extensive range of facial expressions, thereby achieving higher accuracy and robustness in emotion recognition tasks. The choice of ResNet18, complemented by CBAM, offers an efficient and effective means of feature extraction, while the incorporation of Barlow Twins and CARD models enhances the system's learning and predictive capabilities, making it well-suited for the complexities of facial emotion recognition.

5. RESULTS

We presented a comprehensive evaluation of the proposed facial expression recognition (FER) framework, benchmarked against existing methodologies on the FER2013 and AffectNet datasets. The assessments focus on accuracy and the Positive Average Value at Preferred Utility (PAvPU) metrics, with the significance level (α) set at 0.05, as detailed in Tables 1 and 2.

5.1. FER2013 Dataset Results

The performance of our proposed model on the FER2013 dataset is promising, achieving an accuracy of 67.2% and a PAvPU of 66.9%. This demonstrates a considerable improvement over the works of Mao et al. [5], Luo et al. [6], and Wu et al. [24], which reported accuracies of 61.88%, 58.6% (using ResNet18), and 63.25%, respectively. Our model also outperforms the CARD+ResNet18 configuration and its variant with CBAM, which attained accuracies of 61.3% and 64.8%, alongside PAvPU scores of 65.0% and 66.5%, respectively, as shown in Table 1.

Table 1. Quantitative results on FER2013 Dataset

Method	Accuracy (%)	PAvPU ($\alpha = 0.05$)(%)
Mao et al. [5]	61.88	-
Luo et al. [6]	58.6 (ResNet18)	-
Wu wt al. [24]	63.25	-
Proposed	67.2	66.9

5.2. AffectNet Dataset Results

On the AffectNet dataset, our proposed framework secures a leading position with an accuracy of 68.1% and a PAvPU of 67.9%, surpassing the DDAMFN [8], POSTER++ [7], and other referenced methods. Notably, the improvements are particularly significant when compared to the baseline CARD+ResNet18 and its CBAM-enhanced version, which recorded accuracies of 62.3% and 63.9%, respectively. These results, summarized in Table 2, highlight the proposed model's capability to effectively interpret and classify facial expressions with higher precision.

Table 2. Quantitative results on AffectNet Dataset

Method	Accuracy (%)	PAvPU ($\alpha = 0.05$)(%)
DDAMFN [8]	64.25	-
POSTER++ [7]	63.77	-
S2D [9]	63.06	-
Multi-task EfficientNet-B2 [11]	63.03	-
Proposed	68.1	67.9

5.3. Ablation study

We conducted a comprehensive evaluation of our model, analyzing the impact of various building blocks on the results. In initial experiments, we employed a fully supervised learning approach without an attention module, denoted as CARD [13] + ResNet18. Subsequently, in a second experiment, we introduced the attention module, labeled as CARD [13] + ResNet18 + CBAM. Finally, we integrated the complete framework using self-supervised learning. The quantitative results in Table 3 demonstrate the superior performance of our proposed framework.

Table 3. Ablation study of the proposed framework

Method	Accuracy (%)	PAvPU ($\alpha = 0.05$)(%)
FER dataset		
CARD [13] + ResNet18	61.3	65.0
CARD[13] + ResNet18 + CBAM	64.8	66.5
Proposed	67.2	66.9
AffectNet dataset		
CARD [13] + ResNet18	62.3	59.0
CARD [13] + ResNet18 + CBAM	63.9	61.5
Proposed	68.1	67.9

5.4. Comparative Analysis

The comparative analysis reveals the proposed model's superiority in accuracy and PAvPU across both datasets, highlighting its robustness and adaptability in various FER scenarios. The advancements our model introduces set a new benchmark in the field, enhancing the performance of emotion recognition systems in practical applications. The findings of this study not only establish a new standard for FER accuracy but also serve as a foundational reference for future research aiming at the optimization of emotion recognition technologies.

6. CONCLUSION

We presented a Facial Expression Recognition (FER) framework integrating a diffusion-based approach and an attention mechanism. The model, trained efficiently through self-supervised learning on labeled and unlabeled data, exhibited promising performance on the FER2013 and AffectNet datasets, achieving accuracies of 67.2% and 68.1%, respectively. The quantitative results not only outperformed existing state-of-the-art FER models but also showcased the synergistic benefits of combining diffusion-based modeling with self-supervised learning and attention mechanisms within a robust architectural framework. In future, we aim to evaluate our framework on more datasets with complex human emotions. Additionally, we aim to use our framework for recognizing facial expressions of neurological disorder patients.

Acknowledgement

We want to express our gratitude to the Norwegian University of Science and Technology (NTNU) for their invaluable support in preparing the paper. The resources, guidance, financial support, and encouragement provided by the Department of Computer Science have been instrumental in our research and its successful completion.

7. REFERENCES

[1] Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen, "Digiface-1m: 1 million digital face images

for face recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3526–3535.

[2] Muhammad Sajjad, Fath U Min Ullah, Mohib Ullah, Georgia Christodoulou, Faouzi Alaya Cheikh, Mohammad Hijji, Khan Muhammad, and Joel JPC Rodrigues, "A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines," *Alexandria Engineering Journal*, vol. 68, pp. 817–840, 2023.

[3] Zhifeng Wang, Kaihao Zhang, and Ramesh Sankaranarayana, "Lrdif: Diffusion models for under-display camera emotion recognition," *arXiv preprint arXiv:2402.00250*, 2024.

[4] Abdulrahman Alreshidi and Mohib Ullah, "Facial emotion recognition using hybrid features," in *Informatics*. MDPI, 2020, vol. 7, p. 6.

[5] Yuebin Mao, "Optimization of facial expression recgonition on resnet-18 using focal loss and cosface loss," in *2022 International Symposium on Advances in Informatics, Electronics and Education (ISAIEE)*. IEEE, 2022, pp. 161–163.

[6] Jinyu Luo, Zhuocheng Xie, Feiyao Zhu, and Xiaohu Zhu, "Facial expression recognition using machine learning models in fer2013," in *2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC)*. IEEE, 2021, pp. 231–235.

[7] Jiawei Mao, Rui Xu, Xuesong Yin, Yuanqi Chang, Binling Nie, and Aibin Huang, "Poster++: A simpler and stronger facial expression recognition network," *arXiv e-prints*, pp. arXiv–2301, 2023.

[8] Saining Zhang, Yuhang Zhang, Ye Zhang, Yufei Wang, and Zhigang Song, "A dual-direction attention mixed feature network for facial expression recognition," *Electronics*, vol. 12, no. 17, pp. 3595, 2023.

[9] Yin Chen, Jia Li, Shiguang Shan, Meng Wang, and Richang Hong, "From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos," *arXiv e-prints*, pp. arXiv–2312, 2023.

[10] Habib Khan, Mohib Ullah, Fadi Al-Machot, Faouzi Alaya Cheikh, and Muhammad Sajjad, "Deep learning based speech emotion recognition for parkinson patient," *Electronic Imaging*, vol. 35, pp. 298–1, 2023.

[11] Andrey V Savchenko, Lyudmila V Savchenko, and Ilya Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2132–2143, 2022.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[13] Xizewen Han, Huangjie Zheng, and Mingyuan Zhou, "Card: Classification and regression diffusion models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 18100–18115, 2022.

[14] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu, "Attention mechanisms in computer vision: A survey," *Computational visual media*, vol. 8, no. 3, pp. 331–368, 2022.

[15] Keita Mamadou, Mohib Ullah, Øyvind Nordbø, and Faouzi Alaya Cheikh, "Multi-encoder convolution block attention model for binary segmentation," in *2022 International Conference on Frontiers of Information Technology (FIT)*. IEEE, 2022, pp. 183–188.

[16] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 390–391.

[17] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.

[18] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International conference on machine learning*. PMLR, 2021, pp. 12310–12320.

[19] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun 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*. Springer, 2013, pp. 117–124.

[22] Ali Mollahosseini, Behzad Hasani, and Mohammad 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.

[23] Ahmed Halim, Ahmed El-Manfy, Abd El-Rahman Badr, Ali El-Khatib, Mostafa Abd El-Basir, Shehab El-Tabee, Zeyad Alm El-Den, and Asmaa El-Khouly, "Facial expressions analysis to evaluate the level of students' understanding," in *2023 Intelligent Methods, Systems, and Applications (IMSA)*, 2023, pp. 424–429.

[24] Yirui Wu, Lilai Zhang, Zonghua Gu, Hu Lu, and Shaohua Wan, "Edge-ai-driven framework with efficient mobile network design for facial expression recognition," *ACM Transactions on Embedded Computing Systems*, vol. 22, no. 3, pp. 1–17, 2023.