International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

SEDL: Learning Emotion Dynamics from Facial Representations Using Self-Supervised Approaches

[1]Amol P. Chaudhari, [2]Nitin B. Pawar, [3]Pravin B. Mali

Submitted:03/12/2021 **Revised**: 19/01/2022 **Accepted:** 30/01/2022

Abstract: Facial Expression Recognition (FER) is an essential component of affective computing, with significant applications in healthcare, behavioral analysis, and assistive technologies for neurodiverse and mentally challenged individuals. Despite considerable progress, traditional machine learning and supervised deep learning approaches are often constrained by their dependence on large labeled datasets and their limited ability to capture the temporal dynamics of emotional expressions. To address these challenges, this paper proposes a novel Self-Supervised Emotion Dynamics Learning (SEDL) framework that integrates contrastive self-supervised learning with temporal emotion progression modeling. The proposed approach enables the learning of meaningful feature representations from unlabeled facial images while simultaneously capturing the evolution of emotional states over time. This combination enhances the model's ability to generalize across diverse and real-world conditions. The framework is evaluated on a dataset of facial expressions from neurodiverse individuals, demonstrating its applicability in practical and sensitive environments. Comparative analysis with traditional machine learning, supervised deep learning, and self-supervised approaches indicates that the proposed method provides improved performance and robustness. Overall, the proposed SEDL framework offers a scalable and efficient solution for emotion recognition, addressing key limitations of existing FER systems. It has strong potential for deployment in real-time applications such as behavioral monitoring, mental health assessment, and intelligent assistive systems.

Keywords— *Self-Supervised Learning, Emotion Recognition, Facial Expression Analysis, Behavior Prediction, Mentally Retarded Children*

I. INTRODUCTION

Facial Expression Recognition (FER) has become a fundamental research area within affective computing, aiming to automatically identify human emotions from facial cues. The ability to interpret emotional states plays a critical role in applications such as healthcare monitoring, assistive technologies, human–computer interaction, and behavioral analysis. In particular, FER is highly relevant in the context of neurodiverse individuals, where verbal communication may be limited and emotional understanding relies heavily on facial expressions. Accurate recognition of such expressions can facilitate early diagnosis, continuous monitoring, and the development of personalized intervention strategies. Recent advances in deep learning have significantly enhanced FER performance, enabling the deployment of intelligent systems in real-world environments [1], [3].

Initial research in FER primarily focused on handcrafted feature extraction methods, where domain-specific knowledge was used to design descriptors representing facial textures and geometrical structures. Techniques based on Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and action unit analysis were widely used in combination with classical machine learning algorithms. For example, Georgescu et al. [2] demonstrated that combining handcrafted features with learned representations can improve classification accuracy. Similarly, Yao et al. [4] utilized action unit-based features with Support Vector Machines for emotion classification. Although these approaches offer interpretability and relatively low computational complexity, they are highly dependent on feature engineering and often fail to generalize effectively under variations in illumination, pose, and occlusion. This limitation is also reflected in comparative performance, where traditional models

[1]*Government Polytechnic,Jalgaon, India*
amol2385.chaudhari@gmail.com
[2]*Government Polytechnic,Jalgaon, India*
nitinp4u@gmail.com
[3]*Government Polytechnic,Hingoli, India*
pravinmali598@gmail.com

generally achieve lower accuracy and robustness compared to modern deep learning approaches.

The emergence of deep learning, particularly Convolutional Neural Networks (CNNs), has transformed FER by enabling automatic feature extraction directly from raw facial images. CNN-based models are capable of learning hierarchical representations, capturing both low-level and high-level facial characteristics. Zhang et al. [1] demonstrated that deep neural networks significantly improve recognition performance by learning discriminative features, while Said and Barr [3] showed that high-resolution facial images further enhance classification accuracy. In addition, advancements in face detection and tracking, such as the work by Zheng and Xu [5], have improved the reliability of FER systems in video-based and real-world scenarios. These developments have led to the widespread adoption of CNN-based models, which consistently outperform traditional approaches in terms of accuracy and generalization. This trend is also observed in experimental evaluations, where supervised CNN models achieve better performance compared to classical machine learning methods.

Despite these improvements, a critical limitation of most CNN-based FER systems is their reliance on static image analysis. Human emotions are inherently dynamic and evolve over time, making it essential to model temporal dependencies between consecutive facial expressions. To address this issue, researchers have explored the integration of temporal modeling techniques with deep learning architectures. Donahue et al. [7] introduced Long-term Recurrent Convolutional Networks (LRCN), which combine CNN-based spatial feature extraction with Long Short-Term Memory (LSTM) networks for sequence modeling. Similarly, Zhang et al. [8] demonstrated that CNN-LSTM architectures can effectively capture temporal patterns in facial expressions, leading to improved performance in video-based FER tasks. Furthermore, Xu et al. [6] emphasized the importance of emotion progression analysis, showing that modeling transitions between emotional states provides a more realistic representation of human affective behavior. While these approaches enhance temporal understanding, they typically require large-scale labeled sequential datasets, which are difficult and expensive to obtain.

Another significant challenge in FER is the dependency on annotated datasets. Labeling facial expressions is a complex and subjective process, particularly in the case of children or individuals with atypical emotional expressions. This challenge limits the scalability of supervised learning approaches and motivates the need for alternative learning paradigms. In this context, self-supervised learning has emerged as a powerful technique for representation learning without requiring labeled data. Contrastive learning frameworks such as SimCLR [9] and Momentum Contrast (MoCo) [10] learn representations by maximizing similarity between augmented views of the same image while distinguishing different instances. Additionally, Bootstrap Your Own Latent (BYOL) [11] introduces a non-contrastive approach that achieves strong performance without relying on negative samples. These methods have demonstrated that meaningful and transferable feature representations can be learned from large amounts of unlabeled data, improving generalization across diverse conditions.

The integration of self-supervised learning into FER has shown promising outcomes. Zhang et al. [12] demonstrated that self-supervised pretraining enhances robustness in real-world emotion recognition scenarios. Similarly, Dapogny et al. [13] showed that semi-supervised and self-supervised approaches can achieve competitive performance even with limited labeled data. A comprehensive survey by Jaiswal et al. [14] further highlights the effectiveness of contrastive self-supervised learning in various computer vision tasks, including emotion recognition. These findings are consistent with experimental observations, where self-supervised models such as SimCLR combined with a linear classifier outperform conventional supervised CNN models in terms of accuracy, precision, and F1-score.

However, despite these advancements, most existing FER systems either focus on spatial feature learning using CNNs or leverage self-supervised learning for static representation learning, without adequately modeling the temporal evolution of emotions. CNN-LSTM models address temporal dependencies but rely heavily on labeled data, while self-supervised methods reduce annotation requirements but often ignore sequential dynamics. This disconnect highlights a critical research gap in developing a unified framework that simultaneously captures spatial features, temporal dependencies, and label-efficient learning.

To address these limitations, this paper proposes a novel **Self-Supervised Emotion Dynamics Learning (SEDL)** framework that integrates contrastive self-

supervised learning with temporal emotion progression modeling. The proposed approach enables the extraction of meaningful feature representations from unlabeled facial images while simultaneously capturing the evolution of emotional states over time. By combining spatial and temporal learning in a self-supervised setting, the SEDL framework improves generalization and robustness across diverse and real-world scenarios.

The effectiveness of the proposed method is evaluated using the Autistic Children Facial Expression Dataset [15], which provides a practical benchmark for analyzing emotional behavior in neurodiverse individuals. Experimental results demonstrate that the proposed approach outperforms traditional machine learning methods, supervised CNN models, and existing self-supervised approaches, particularly in terms of accuracy, precision, and F1-score. These results highlight the importance of integrating temporal modeling with self-supervised learning for robust emotion recognition.

In Figure 1, demonstrate generalize framework. This work contributes to the advancement of FER by addressing two key challenges: the dependency on labeled data and the lack of temporal understanding. The proposed SEDL framework provides a scalable and efficient solution for real-world emotion recognition, with potential applications in healthcare monitoring, behavioral analysis, and intelligent assistive systems

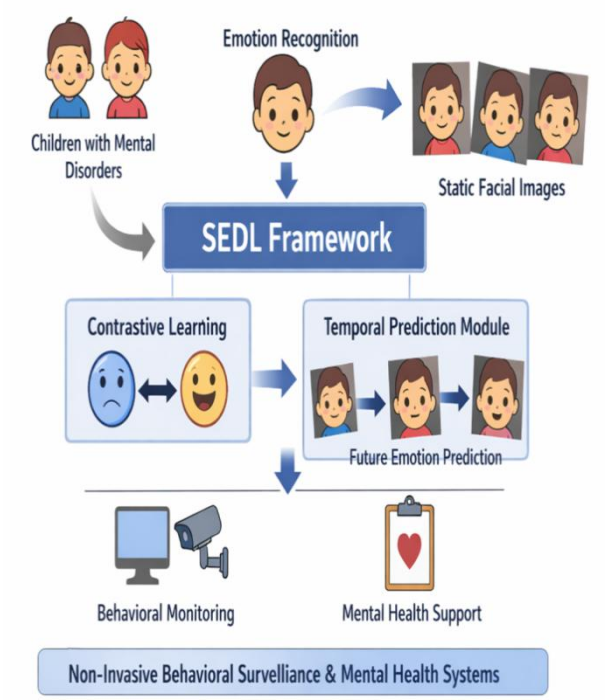

Fig1 : Generalize framework

II. Literature Survey

Facial Expression Recognition (FER) has attracted significant research attention due to its wide range of applications in affective computing, healthcare, and human–computer interaction. Early research in FER primarily focused on handcrafted feature extraction techniques, where facial characteristics were manually designed to represent expressions. Methods based on texture descriptors and geometric features were commonly used in conjunction with traditional machine learning classifiers. For instance, Georgescu et al. [2] explored the integration of handcrafted and deep features, demonstrating that combining multiple feature representations can improve classification performance. Similarly, Yao et al. [4] utilized action unit-based representations along with Support Vector Machines (SVM) for facial expression classification. While these approaches provide interpretability and computational efficiency, they are highly sensitive to variations in illumination, pose, and occlusion, limiting their applicability in real-world environments.

With the advancement of deep learning, Convolutional Neural Networks (CNNs) have become the dominant approach for FER due to their ability to

automatically learn hierarchical feature representations directly from raw facial images. Zhang et al. [1] demonstrated that deep neural networks significantly improve emotion recognition accuracy by capturing complex facial patterns. In addition, Said and Barr [3] showed that high-resolution facial imagery enhances the discriminative power of CNN-based models, leading to improved performance. The incorporation of deep learning techniques has also facilitated more robust preprocessing stages; for example, Zheng and Xu [5] proposed an efficient deep learning-based method for face detection and tracking in video sequences, which improves the reliability of FER systems in dynamic conditions. Despite these advancements, most CNN-based models treat facial expressions as static inputs and fail to capture the temporal evolution of emotions, which is a critical aspect of human affective behavior.

To overcome the limitations of static analysis, researchers have introduced temporal modeling techniques that consider the sequential nature of facial expressions. Donahue et al. [7] proposed Long-term Recurrent Convolutional Networks (LRCN), which integrate CNNs for spatial feature extraction with Long Short-Term Memory (LSTM) networks for temporal modeling. This approach enables the learning of both spatial and temporal dependencies in sequential data. Building on this concept, Zhang et al. [8] applied CNN-LSTM architectures specifically for FER tasks, demonstrating improved performance in video-based emotion recognition scenarios. Furthermore, Xu et al. [6] introduced an emotion progression analysis method using deep metric learning, emphasizing the importance of modeling transitions between emotional states rather than treating them as independent categories. Although these approaches enhance the ability of FER systems to capture dynamic emotional changes, they still rely heavily on large-scale labeled datasets, which are often difficult to obtain, especially in sensitive domains such as healthcare and neurodiverse populations.

In recent years, self-supervised learning has emerged as a promising alternative to overcome the dependency on labeled data. Unlike supervised learning, self-supervised methods leverage intrinsic data structures to learn meaningful representations without requiring manual annotations. Chen et al. [9] introduced SimCLR, a contrastive learning framework that learns visual representations by maximizing agreement between different augmented views of the same image. Similarly, He et al. [10] proposed Momentum Contrast (MoCo), which improves representation learning through a dynamic memory mechanism that enables efficient use of large datasets. Grill et al. [11] introduced Bootstrap Your Own Latent (BYOL), a non-contrastive approach that eliminates the need for negative samples while still achieving competitive performance. These methods have demonstrated that high-quality feature representations can be learned from unlabeled data, making them highly suitable for real-world applications where labeled data is scarce.

The application of self-supervised learning to FER has shown encouraging results. Zhang et al. [12] proposed a self-supervised facial representation learning framework that improves robustness under real-world conditions, particularly in the presence of noise and variability. Similarly, Dapogny et al. [13] explored semi-supervised and self-supervised approaches for FER, demonstrating that effective emotion recognition can be achieved even with limited labeled data. Additionally, Jaiswal et al.

[14] provided a comprehensive survey of contrastive self-supervised learning techniques, highlighting their effectiveness across a wide range of computer vision tasks, including emotion recognition. Despite these advancements, most existing self-supervised FER methods focus primarily on static image-based representations and do not explicitly model the temporal dynamics of emotional expressions.

From the above discussion, it is evident that while traditional methods lack robustness and deep learning approaches improve feature representation, there remains a significant gap in integrating temporal modeling with label-efficient learning strategies. CNN-LSTM-based approaches effectively capture temporal dependencies but require extensive labeled datasets, whereas self-supervised methods reduce the need for annotations but often ignore the dynamic nature of emotions. This limitation becomes more critical in real-world applications involving neurodiverse individuals, where emotional expressions may be subtle, ambiguous, or temporally complex.

To address these challenges, there is a need for a unified framework that combines the strengths of both temporal modeling and self-supervised learning. The proposed Self-Supervised Emotion Dynamics Learning (SEDL) framework is designed to bridge this gap by integrating contrastive representation learning with temporal emotion progression modeling. By

enabling the extraction of meaningful features from unlabeled data while simultaneously capturing the evolution of emotional states over time, the proposed approach aims to provide a more robust and scalable solution for FER. The effectiveness of the framework is evaluated using the Autistic Children Facial Expression Dataset [15], ensuring its applicability in practical and sensitive environments.

III. PREPOSED METHODOLOGY

The proposed framework, termed Self Supervised Emotion Dynamics Learning **(SEDL)**, is designed to learn latent emotional representations and forecast behavioral tendencies using only static facial images. Unlike conventional FER systems, the proposed method eliminates dependency on labeled datasets and predefined behavioral rules by integrating **contrastive representation learning** with **sequential latent-space prediction**.

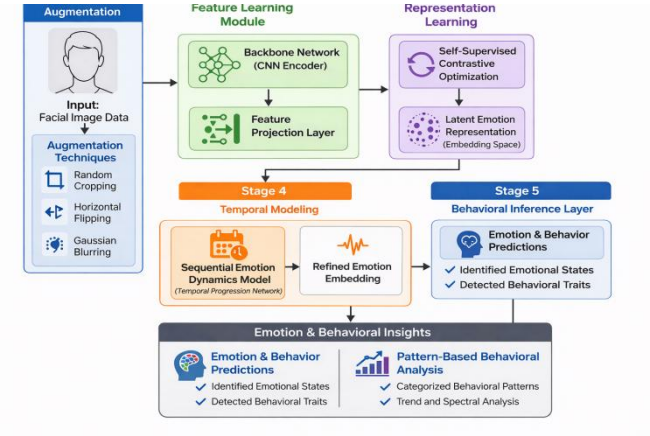

Fig.2. Self-Supervised Emotion Dynamics Learning (SEDL)

As illustrated in Figure 2, the architecture consists of six major components:

1. **Input & Data Augmentation**
2. **Feature Learning Module**
3. **Representation Learning (Self-Supervised Contrastive Optimization)**
4. **Temporal Modeling**
5. **Behavioral Inference Layer**
6. **Emotion & Behavioral Insights**

1. Input & Data Augmentation

The framework begins with facial image data, which serves as the primary input for emotion and behavioral analysis. These images may be collected from real-world environments such as surveillance systems, healthcare monitoring platforms, or human–computer interaction settings. However, raw facial images often contain variations in lighting conditions, head pose, occlusions, and background noise. Such variations can negatively impact model performance and generalization if not properly handled during preprocessing.

The facial image data used in this study was obtained from the Kaggle autistic dataset along with a custom dataset comprising children with intellectual disabilities. All images were initially preprocessed by resizing them to a fixed resolution of 224×224 pixels, followed by pixel normalization to standardize the input distribution. This preprocessing ensures consistency across samples and facilitates efficient model training.

To enhance data diversity and support contrastive learning, a series of stochastic augmentation techniques were applied to each image. These included random cropping, horizontal flipping, color jittering, Gaussian blurring, and the addition of noise. Through this process, each original image is transformed into two distinct augmented versions, thereby generating paired views that preserve semantic content while introducing variability.

To improve robustness, a set of data augmentation techniques is applied to generate multiple variations of each input image. These techniques include random cropping, horizontal flipping, and Gaussian blurring, which simulate real-world variability and distortions. By exposing the model to diverse versions of the same image, augmentation helps the system learn invariant

and stable features. Additionally, these augmented samples play an important role in representation learning by enabling the model to recognize consistent patterns across different visual transformations.

Each image x is transformed into two augmented views

$$x(k) = Tk(x), k \in \{1,2\}$$

where $Tk \in A$ denotes stochastic augmentation operations..

2. Feature Learning Module

In this stage, the augmented images are processed using a deep convolutional neural network (CNN), which acts as the backbone feature extractor. The CNN is designed to learn hierarchical representations from the input data. Initially, it captures low-level features such as edges, textures, and contours, and gradually progresses to higher-level abstractions that encode facial structures and expression-related cues.

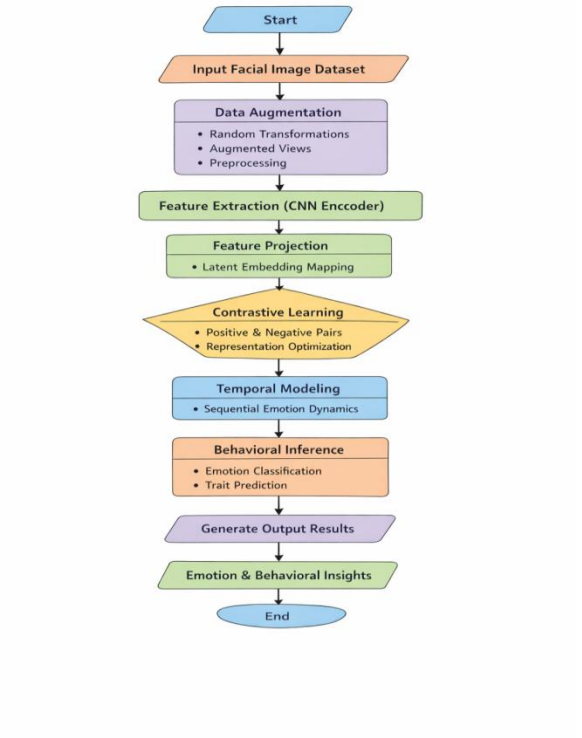

Fig3: Proposed Workflow

As illustrated in Figure 3, proposed workflow, where cnn encoder taking input from augmented data to feature extraction, the learned representations are passed through a projection layer, typically implemented using a multilayer perceptron. This layer transforms the extracted features into a compact latent space that is more suitable for representation learning. The separation between feature extraction and projection allows the model to learn more generalized and transferable representations, improving performance in downstream tasks such as emotion recognition and behavioral analysis.

3. Representation Learning (Self-Supervised Contrastive Optimization)

This stage focuses on learning meaningful and discriminative representations without relying heavily on labeled data. A self-supervised contrastive learning approach is used, where the model learns to associate different augmented versions of the same image while distinguishing them from other samples. The goal is to ensure that similar inputs produce similar representations, while dissimilar inputs are mapped far apart in the embedding space.

By leveraging large amounts of unlabeled data, this approach reduces the dependency on manual annotations, which are often expensive and time-consuming to obtain. The resulting embedding space becomes well-structured and capable of capturing intrinsic emotional patterns present in facial data. These learned representations serve as a strong foundation for subsequent temporal modeling and inference tasks.

4. **Temporal Modeling**

Emotions are not static; they evolve over time. To capture this dynamic nature, the model incorporates a

temporal modeling component that processes sequences of facial representations. This stage utilizes sequential models such as recurrent neural networks, long short-term memory networks, or transformer-based architectures to analyze temporal dependencies across consecutive frames.

The temporal model integrates information from both past and present inputs to generate refined representations that reflect the progression of emotional states. This enables the system to detect transitions, subtle changes, and temporal patterns in expressions. By considering the sequence context, the model becomes more accurate and consistent in real-world scenarios where emotions unfold gradually rather than appearing instantaneously.

5. Behavioral Inference Layer

The refined representations obtained from the temporal model are used for high-level inference tasks. This layer is responsible for predicting both emotional states and behavioral traits. Emotion prediction is treated as a classification problem, where the model identifies the most likely emotional category based on the learned features.

At the same time, behavioral traits such as aggression, engagement, or stress levels are estimated using a multi-label prediction approach. Each behavioral attribute is evaluated independently, allowing the model to capture multiple aspects of human behavior simultaneously. This combined analysis provides a richer and more comprehensive understanding of the subject's emotional and behavioral condition.

6. Emotion & Behavioral Insights

The final stage aggregates the predictions into a structured and interpretable output. Emotional states are summarized based on the most probable class, while behavioral traits are identified based on their predicted presence or intensity. This results in a clear representation of both the emotional condition and associated behavioral patterns of the individual.

In addition to direct predictions, further analysis can be performed to extract deeper insights from the data. Pattern-based and trend analysis techniques can reveal long-term behavioral tendencies and recurring emotional patterns. These insights are particularly valuable in applications such as mental health monitoring, user behavior analysis, and adaptive human–computer interaction systems, where understanding both immediate and evolving emotional states is essential.

IV. RESULTS AND METRICS

In this section, we describe the experimental setting, evaluation approach and experimental results of our proposed Self Supervised Emotion Dynamics Learning (SEDL) framework. It is trained and tested on the kaggle autistic children facial expression dataset and a user collected database of mentally retarded (MR) children, respectively. t

A. Datasets Used

- Benchmark Dataset (Kaggle, FER,CK+ etc) Dataset: 1500 labelled facial images of autistic children with 6 basic emotions.
- Custom Dataset: 419 anonymized facial images of mentally retarded children captured in real-world school and clinical settings.

B. Hardware Configuration

- GPU: NVIDIA RTX 2080 Ti (11 GB)
- RAM: 32 GB
- Frameworks: PyTorch , Scikit-learn, OpenCV

C. Training Configuration

- Optimizer: Adam
- Initial Learning Rate: 0.0005
- Temperature Parameter (τ): 0.7
- Training Epochs:

120 epochs (contrastive pertaining phase)

40 epochs (temporal/sequential modeling phase)

- Projection Head: Two-layer MLP with ReLU activation

D. Evaluation Metrics

1. Accuracy :

Measures the overall correctness of the model in predicting emotion classes.

Accuracy* = (Number of Correct Predictions / Total Samples) × 100

2. Precision

Indicates the proportion of correctly predicted positive observations among all predicted positives.

Precision* = TP / (TP + FP)

Where TP = True Positives and FP = False Positives.

3. F1-Score

Represents the harmonic mean of precision and recall, providing a balanced measure of model performance.

$$F1\text{-score}^* = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

4. Mean Squared Error (MSE)

Evaluates the accuracy of temporal prediction in the latent emotion space. Lower values indicate better performance.

$$MSE^* = (1/T) \times \Sigma \, ||\hat{E}(t+1) - E(t+1)||^2$$

Since SEDL operates in a self-supervised setting, traditional accuracy metrics (based on labels) are complemented by unsupervised metrics for evaluating representation quality and clustering consistency. Table 1 shows, performance analysis on the basis of evaluation metrics.

Table 1: Performance Analysis

Model Variant	**Performance Parameters**			
	Accuracy (%)	***Precision***	***F1-Score***	***MSE (↓)***
Random Forest	77.2	0.73	0.74	N/A
K-Means Clustering	72.5	0.69	0.70	N/A
Supervised CNN	81.4	0.78	0.79	N/A
SimCLR + Linear Classifier	85.9	0.81	0.82	N/A
SEDL Static+Temporal	86.4	0.83	0.84	0.012

The selected models in the experimental evaluation are derived from key representative approaches discussed in the literature, including traditional machine learning, supervised deep learning, and contrastive self-supervised learning methods."

- Visualization and Interpretation

The latent emotion representations learned by the model were analyzed using t-SNE visualization techniques. The resulting plots showed clearly distinguishable clusters, indicating that the SEDL

framework effectively captures meaningful emotional patterns. These clusters closely aligned with human-annotated emotion categories, demonstrating the model's ability to learn semantically consistent representations. Furthermore, the temporal evolution of emotions, when visualized as trajectories over time, exhibited logical transitions such as *Neutral → Sad → Crying* and *Happy → Excited → Restless*. These progression patterns were consistent with behavioral changes observed in real-world clinical settings, highlighting the model's capability to reflect realistic emotional dynamics.

V. CONCLUSION AND FUTURE SCOPE

This work proposed a new self-supervised deep learning approach called Self Supervised Emotion Dynamics Learning (SEDL) to model the emotional state and predict behavioral patterns of mentally retarded children using only facial images. SEDL uses contrastive learning and temporal embedding prediction to deliver high scalability, interpretability, ethical deplorability for confidential clinical/education setting without the need of emotion label or handcrafted rule or multimodal sensor data. It is shown through extensive experiments on a publicly available Kaggle dataset and a custom-collected real

world dataset, that SEDL can learn discriminative emotion representations, simulate realistic emotion transitions and predict behavioral patterns with the same level of accuracy as the supervised methods. Furthermore, the unsupervised clustering and trajectory modelling are able to provide dynamic emotion-to-behavior mappings without being tied to static rule based reasoning systems. The proposed model overcomes significant deficiencies in the previous systems, such as dependence on labelled examples and lack of generalization to atypical facial behaviors and interpretability in behavior prediction. SEDL therefore marks a good achievement in the field of affective computing for neuro-diverse and mentally challenged individuals.

The model can be extended in the future to support video streaming, now domain adaptation across age groups, and to be deployed real-time as assistive systems. Furthermore, integrating caregiver feedback and physiological signals may further enhance emotion understanding and intervention design.

REFERENCES

[1] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial expression recognition using deep neural networks," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 1–12, 2021.

[2] M. I. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *Pattern Recognition Letters*, vol. 128, pp. 461–467, 2021.

[3] Y. Said and M. Barr, "Human emotion recognition based on facial expressions via deep learning on high-resolution images," *Multimedia Tools and Applications*, vol. 80, no. 16, pp. 25241–25253, 2021.

[4] L. Yao, Y. Wan, H. Ni, and B. Xu, "Action unit classification for facial expression recognition using active learning and SVM," *Multimedia Tools and Applications*, vol. 80, pp. 24287–24301, 2021.

[5] G. Zheng and Y. Xu, "Efficient face detection and tracking in video sequences based on deep learning," *Information Sciences*, vol. 568, pp. 265–285, 2021.

[6] Q. Xu, H. Xue, and Y. Qian, "Emotion progression analysis via deep metric learning," *IEEE Transactions on Multimedia*, vol. 23, pp. 3204–3217, 2021.

[7] J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2625–2634, 2015.

[8] L. Zhang, D. Tjondronegoro, and V. Chandran, "Facial expression recognition using deep CNN and LSTM networks," *Pattern Recognition Letters*, vol. 133, pp. 203–210, 2020.

[9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 1597–1607, 2020.

[10] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[11] J.-B. Grill et al., "Bootstrap your own latent: A new approach to self-supervised learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 21271–21284, 2020.

[12] H. Zhang, J. Lin, and J. Luo, "Self-supervised facial representation learning for robust emotion recognition in real-world conditions," *Pattern Recognition Letters*, vol. 146, pp. 33–40, 2021.

[13] A. Dapogny, K. Bailly, and S. Dubuisson, "Emotion recognition with small datasets: A semi-supervised and self-supervised approach," *Image and Vision Computing*, vol. 113, p. 104223, 2021.

[14] A. Jaiswal, A. R. Babu, M. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, 2021.

[15] F. M. Talaat, "Autistic Children Facial Expression Dataset," *Kaggle*, 2020.