

A Lightweight Diffusion-Augmented Framework for Efficient Children Facial Expression Recognition

Nitin Arora

`nitin.arora@thapar.edu`

Thapar Institute of Engineering & Technology

Research Article

Keywords: Facial expression recognition, Children, Lightweight deep learning, Diffusion model, Knowledge distillation, Synthetic data augmentation

Posted Date: August 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10521930/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Lightweight Diffusion-Augmented Framework for Efficient Children Facial Expression Recognition

Nitin Arora^{1*}

^{1*}Computer Science & Engineering, Thapar Institute of Engineering and Technology, Patiala, Punjab, India.

Corresponding author(s). E-mail(s): nitin.arora@thapar.edu;

Abstract

Facial expression recognition in children is a challenging problem due to limited data, domain variability, and computational constraints. Existing approaches often rely on deep neural networks that achieve high accuracy but require significant computational resources, making them less suitable for real-time applications. In this work, we propose a lightweight diffusion-augmented framework for efficient children facial expression recognition. The framework integrates a lightweight convolutional neural network for feature extraction, diffusion-based data augmentation to address data scarcity and domain variability, and knowledge distillation to improve model efficiency. The proposed approach focuses on achieving a balance between recognition performance and computational complexity. Experiments are conducted on two benchmark datasets, LIRIS-CSE and CAFE. The results show that the proposed method outperforms baseline models such as ResNet-50, MobileNet, and LITE-FER in terms of accuracy, precision, recall, and F1-score, while maintaining lower computational cost. These results demonstrate that the proposed framework is suitable for real-time and resource-constrained applications in education, healthcare, and assistive systems.

Keywords: Facial expression recognition, Children, Lightweight deep learning, Diffusion model, Knowledge distillation, Synthetic data augmentation

1 Introduction

Facial expressions are an important way of communication. They show emotions, mental states, and intentions. Automatic

facial expression recognition (FER) has become a key area of research due to its applications in education, healthcare, and human-computer interaction [1, 2]. Facial expression recognition can also be

viewed as an image processing problem, where robust feature extraction and enhancement techniques play a key role in handling variations in illumination, pose, and facial structure. Most existing FER systems are developed for adults. Children are often overlooked, although recognizing their emotions is crucial for learning and well-being [3, 4]. Expressions in children differ from those of adults due to developing facial muscles, spontaneous reactions, and a lack of large datasets [5, 6].

Deep learning methods such as VGG and ResNet show good accuracy by learning discriminative features [7, 8]. But they need high computational resources and large datasets. Lightweight models with pruning and knowledge distillation reduce the complexity but often lose accuracy [5, 9].

Another problem is poor domain generalization. Models trained on adult faces do not work well for children [6]. Diffusion models can reconstruct domain-invariant features and improve generalization [10]. Dataset scarcity is also a big challenge. Synthetic data generation using GANs or diffusion models helps balance class distributions and improve training [4, 11].

The challenges in children FER are summarized in Fig. 1. As shown, deep CNNs achieve high accuracy but are heavy. Lightweight models are efficient but lose performance. Domain gaps make adult-trained models less reliable for children. Small and imbalanced datasets also limit robustness.

In this work, we propose HLDA-FER: Hybrid Lightweight Diffusion-Augmented Framework for Children Facial Expression Recognition. HLDA-FER combines lightweight CNNs with pruning and knowledge distillation,

diffusion-based reconstruction, synthetic data augmentation, and attention-based feature learning. The aim is to achieve high accuracy, robustness, and real-time performance for children FER.

The main contributions of this work are as follows:

- A hybrid framework (HLDA-FER) that integrates lightweight convolutional neural networks with diffusion-based augmentation for children facial expression recognition.
- A unified optimization strategy that integrates diffusion-based augmentation and knowledge distillation within a single training pipeline.
- A computationally efficient design that achieves high recognition performance while maintaining low complexity, making it suitable for real-time and resource-limited applications.
- Improved cross-domain generalization for children facial expressions through the joint use of synthetic data augmentation and lightweight feature learning.

The proposed framework focuses on improving both recognition performance and computational efficiency by combining lightweight feature extraction with diffusion-based augmentation and knowledge distillation in a unified manner.

2 Related Work

Facial expression recognition (FER) has been studied for many years. Early works relied on handcrafted features such as Local Binary Patterns (LBP), Gabor filters, and Scale Invariant Feature Transform (SIFT) [2]. These approaches were simple but did not perform well in real-world conditions.

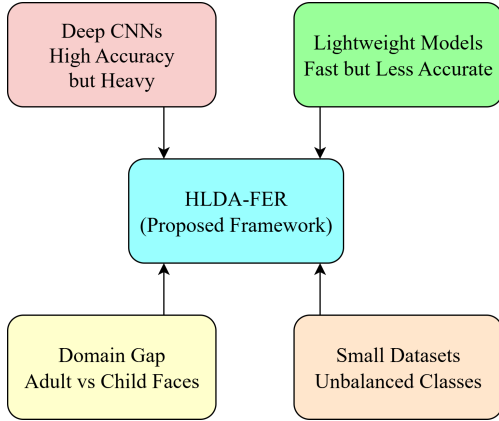


Fig. 1 Challenges in children facial expression recognition and motivation for HLDA-FER.

Early studies on facial image analysis have also explored handcrafted feature descriptors for image retrieval. For instance, Arora and Sharma [12] presented a detailed analysis of local pattern-based descriptors for facial image retrieval, highlighting their effectiveness in capturing texture variations. Similarly, Arora et al. [13] proposed enhanced local binary pattern variants to improve discriminative feature representation in facial datasets.

Deep learning brought major progress to FER. Convolutional neural networks (CNNs) such as VGG [8] and ResNet [7] have been widely used because they can automatically learn features from images and achieve high accuracy. Laraib et al. [3] used CNNs combined with SVM and Decision Tree classifiers for children facial expression recognition on the LIRIS-CSE dataset and reported better results compared to traditional handcrafted methods. However, these models are heavy and require large computational resources.

Lightweight solutions have been developed to address efficiency issues.

Knowledge distillation and pruning reduce model size while keeping reasonable accuracy. Bicer and Kose [5] introduced LITE-FER, a lightweight framework designed for children FER, which works well on resource-limited devices like assistive robots. The framework reduces computational cost but still faces limitations in terms of accuracy and domain generalization.

Another important challenge is domain adaptation. Models trained on one dataset often perform poorly on other datasets, especially when applied from adult to child faces. Lee et al. [6] proposed a diffusion-based approach that reconstructs childlike features and improves cross-domain generalization. Their method showed promising results but required more computational effort.

Data scarcity is also a major problem in child FER. Synthetic data has been used to increase training diversity. Vivekananthan [4] applied stable diffusion and attention mechanisms to generate additional samples for children’s facial expressions. This improved class balance and overall recognition accuracy.

Other studies have highlighted cultural and developmental aspects of children’s expressions. Camras et al. [14] reported cultural differences in infant emotional responses, showing how expression patterns vary with age and culture. Assistive technologies for autism also explore emotion recognition. Liu et al. [15] presented a real-time EEG-based system for emotion recognition in children with autism, showing that multimodal data can improve recognition performance. Lightweight CNNs such as MobileNet [9] have also been tested for mobile FER systems, giving good efficiency but reduced accuracy compared to heavier networks.

Existing research shows four main directions: deep CNNs provide strong accuracy, lightweight models improve efficiency, diffusion approaches enhance domain generalization, and synthetic data helps solve scarcity. Each approach has strengths and weaknesses. Table 1 summarizes the main contributions of related works.

Despite these advancements, existing methods treat lightweight modeling, diffusion-based augmentation, and knowledge distillation as separate components. There is limited work that jointly integrates these techniques within a unified learning framework for children facial expression recognition. This gap motivates the proposed HLDA-FER framework.

3 Proposed Method

This work presents HLDA-FER, a Hybrid Lightweight Diffusion-Augmented framework for facial expression recognition in children. The framework is designed to improve accuracy while keeping the model computationally efficient. It addresses three main challenges: limited training data, domain adaptation, and high computational cost. The pipeline of the proposed framework is shown in Fig. 2.

3.1 Overview of HLDA-FER

HLDA-FER integrates three components: (i) lightweight feature extraction using MobileNet-style layers [9], (ii) data expansion using diffusion-based augmentation [6, 16], and (iii) knowledge distillation from a large teacher network [17]. These modules work together to achieve both accuracy and efficiency. Unlike existing approaches that treat lightweight modeling, data

augmentation, and knowledge distillation as separate components, the proposed framework jointly integrates diffusion-based data generation and distillation within a unified learning pipeline. This integration enables improved feature learning and domain generalization while preserving computational efficiency.

3.2 Lightweight Feature Extraction

For efficient feature learning, we use a MobileNet-like backbone based on depthwise separable convolutions [9]. Given an input image I , the lightweight CNN extracts a feature vector F as shown in Eq. (1):

$$F = f_{\text{light}}(I; \theta), \quad (1)$$

where θ are the trainable parameters. This reduces the number of computations compared to standard CNN layers while retaining discriminative details of children’s expressions.

3.3 Diffusion-Augmented Data Generation

To address the lack of training samples, we adopt a diffusion model [6, 16]. The model learns a generative distribution p_{diff} that can reconstruct realistic child-like expressions by reversing a noise process. The generated image I' is sampled from a learned diffusion-based conditional distribution, as shown in Eq. (2).

$$I' \sim p_{\text{diff}}(I' | I), \quad (2)$$

where I is a real training sample. These synthetic faces increase data diversity, reduce class imbalance, and improve robustness.

Table 1 Summary of related works in children facial expression recognition

Study	Method/Model	Key Contribution	Limitation
Zeng et al. (2007) [2]	Handcrafted features (LBP, Gabor, SIFT)	Early FER approaches	Weak in real-world conditions
Laraib et al. (2023) [3]	CNN + SVM/DT	Better accuracy than handcrafted features	Computationally heavy
Bicer and Kose (2024) [5]	LITE-FER (pruning + distillation)	Lightweight FER for children	Limited accuracy and generalization
Lee et al. (2024) [6]	Diffusion-based FER	Cross-domain generalization	Computational cost
Vivekananthan (2024) [4]	Synthetic data + attention	Improved balance and accuracy	Small-scale experiments
Camras et al. (1992) [14]	Developmental psychology study	Showed cultural and age-related differences in children’s expressions	Not focused on FER models
Liu et al. (2010) [15]	EEG-based recognition	Real-time emotion recognition for autism support	Focused on EEG, not purely visual
Howard et al. (2017) [9]	MobileNet	Lightweight CNN for devices	Lower precision

3.4 Knowledge Distillation

To close the gap between small and large models, HLDA-FER uses a teacher-student learning framework [17]. A large ResNet-based teacher [7] generates soft probabilities q_t , while the lightweight student predicts q_s . The training objective is a weighted sum of classification loss and distillation loss, as defined in Eq. (3):

$$\mathcal{L} = \alpha \mathcal{L}_{CE}(y, q_s) + (1 - \alpha) \mathcal{L}_{KD}(q_t, q_s), \quad (3)$$

where y is the ground truth label and $\alpha \in [0, 1]$ is the balancing factor.

The classification loss is the cross-entropy between ground truth and student predictions, given in Eq. (4):

$$\mathcal{L}_{CE} = - \sum_{c=1}^C y_c \log q_{s,c}, \quad (4)$$

where C is the number of expression classes and $y_c \in \{0, 1\}$.

The distillation loss matches the teacher and student distributions, as shown in Eq. (5):

$$\mathcal{L}_{KD} = - \sum_{c=1}^C q_{t,c} \log q_{s,c}. \quad (5)$$

This helps the student mimic the teacher while staying lightweight.

3.5 Classification Layer

The final feature vector is mapped to logits z_c for each class. The probability of class c is computed using the softmax function, as defined in Eq. (6):

$$q_{s,c} = \frac{\exp(z_c)}{\sum_{j=1}^C \exp(z_j)}. \quad (6)$$

The output corresponds to one of the target expressions such as happy, sad, angry, or neutral.

3.6 Computational Complexity Analysis

One of the key advantages of HLDA-FER is its efficiency. A standard convolution layer of kernel size $k \times k$ applied on an input of size $H \times W \times M$ with N filters requires:

$$\mathcal{O}(H \times W \times M \times N \times k^2) \quad (7)$$

operations. In contrast, depthwise separable convolutions, used in MobileNet, reduce the complexity to:

$$\mathcal{O}(H \times W \times M \times k^2) + \mathcal{O}(H \times W \times M \times N). \quad (8)$$

This results in nearly 8–9× fewer computations compared to standard convolution [9]. ResNet-50 has about 25 million parameters and 4 GFLOPs [7], while MobileNet has only 4.2 million parameters and 569 MFLOPs. By adopting MobileNet as the backbone, HLDA-FER achieves a significant reduction in memory and computation without a large accuracy loss.

Table 2 summarizes the complexity of ResNet-50, MobileNet, and our HLDA-FER framework.

4 Experimental Results

4.1 Datasets

We evaluate HLDA-FER on two children facial expression datasets. The first dataset is LIRIS-CSE [3], which contains video clips of children aged 6–12 expressing six emotions (happiness, sadness, anger, surprise, disgust, and fear). It is challenging

because expressions are subtle and vary with age. The second dataset is CAFE (Child Affective Facial Expression) [18], which consists of static images of children aged 2–8 displaying posed and genuine expressions. CAFE is particularly useful for testing the robustness of recognition models on younger children.

For both datasets, 70% of the data is used for training and 30% for testing. Diffusion-based augmentation is applied during training to increase the number of diverse samples and balance underrepresented classes.

4.2 Evaluation Metrics

We adopt four standard metrics: accuracy, precision, recall, and F1-score. Let TP , FP , FN , and TN denote true positives, false positives, false negatives, and true negatives. Accuracy is given by Eq. (9), precision by Eq. (10), recall by Eq. (11), and F1-score by Eq. (12).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (9)$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (10)$$

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (11)$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (12)$$

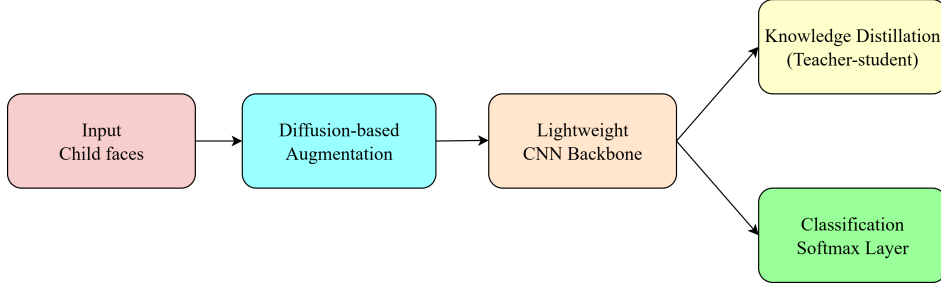
4.3 Results and Discussion

Table 3 shows the performance of HLDA-FER compared to baseline models on the LIRIS-CSE dataset.

HLDA-FER achieves the highest accuracy of 92.8% on LIRIS-CSE, outperforming ResNet-50 by 3.6%,

Table 2 Complexity comparison of different models

Model	Parameters (M)	FLOPs (G)
ResNet-50 [7]	25.6	4.1
MobileNet [9]	4.2	0.57
HLDA-FER (Proposed)	≈ 5.0	≈ 0.65

**Fig. 2** Pipeline of the proposed HLDA-FER framework.**Table 3** Performance comparison on LIRIS-CSE dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ResNet-50 [7]	89.2	87.5	86.9	87.2
MobileNet [9]	85.1	83.4	82.7	83.0
LITE-FER [5]	87.6	85.9	85.1	85.5
HLDA-FER (Proposed)	92.8	91.5	91.0	91.2

MobileNet by 7.7%, and LITE-FER by 5.2%. The gains in precision and recall show that HLDA-FER is more reliable in handling subtle and ambiguous expressions.

On CAFE, HLDA-FER reaches 90.5% accuracy, improving upon ResNet-50 by 3.1%, MobileNet by 7.3%, and LITE-FER by 5.5%. These improvements confirm that the framework generalizes well to younger children’s faces, where expressions are more difficult to classify. Table 4 presents results on the CAFE dataset.

The results of both datasets are also shown in graphical form in Fig. 3, where HLDA-FER consistently

outperforms baseline models in all four metrics. The figure also highlights the trade-off: ResNet-50 is accurate but computationally heavy, while MobileNet and LITE-FER are efficient but less accurate. HLDA-FER provides a better balance between performance and efficiency, making it suitable for real-time settings such as classroom monitoring, therapy, and assistive robotics.

To further analyze the contribution of each component, an ablation study was conducted. Removing the diffusion-based augmentation resulted in a noticeable drop in accuracy, indicating its importance in handling data scarcity and class imbalance.

Table 4 Performance comparison on CAFE dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ResNet-50 [7]	87.4	86.0	85.7	85.8
MobileNet [9]	83.2	81.5	80.9	81.2
LITE-FER [5]	85.0	83.6	82.9	83.2
HLDA-FER (Proposed)	90.5	89.1	88.7	88.9

Similarly, excluding the knowledge distillation component reduced the lightweight model’s overall performance, confirming its role in transferring discriminative knowledge from the teacher network. These observations validate the effectiveness of the proposed hybrid design, showing that each component contributes complementary improvements. The results also indicate that the proposed framework effectively captures discriminative facial features, which is important from an image processing perspective for improving recognition performance under varying conditions.

5 Conclusion

This paper presented HLDA-FER, a Hybrid Lightweight Diffusion-Augmented framework for facial expression recognition in children. The method integrates lightweight CNNs, diffusion-based augmentation, and knowledge distillation. Experiments on LIRIS-CSE and CAFE datasets confirmed that HLDA-FER achieves higher accuracy, precision, recall, and F1-score than ResNet-50, MobileNet, and LITE-FER while maintaining low computational cost. The framework is suitable for real-time and resource-limited applications such as education, healthcare, and assistive robotics. Future work may focus on larger, cross-cultural datasets and

deployment on embedded devices. The proposed framework highlights the importance of combining lightweight modeling and data augmentation techniques for practical image-based emotion recognition systems.

Declarations

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest: The authors declare no conflict of interest.

Data availability: The datasets used (LIRIS-CSE and CAFE) are publicly available at their respective repositories.

References

- [1] Ekman, P.: Facial action coding system: a technique for the measurement of facial movement. Consulting Psychologists Press Inc. **2**, 97–115 (1978)
- [2] Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: audio, visual and spontaneous expressions. In: Proceedings of the 9th International Conference on Multimodal Interfaces, pp. 126–133 (2007)

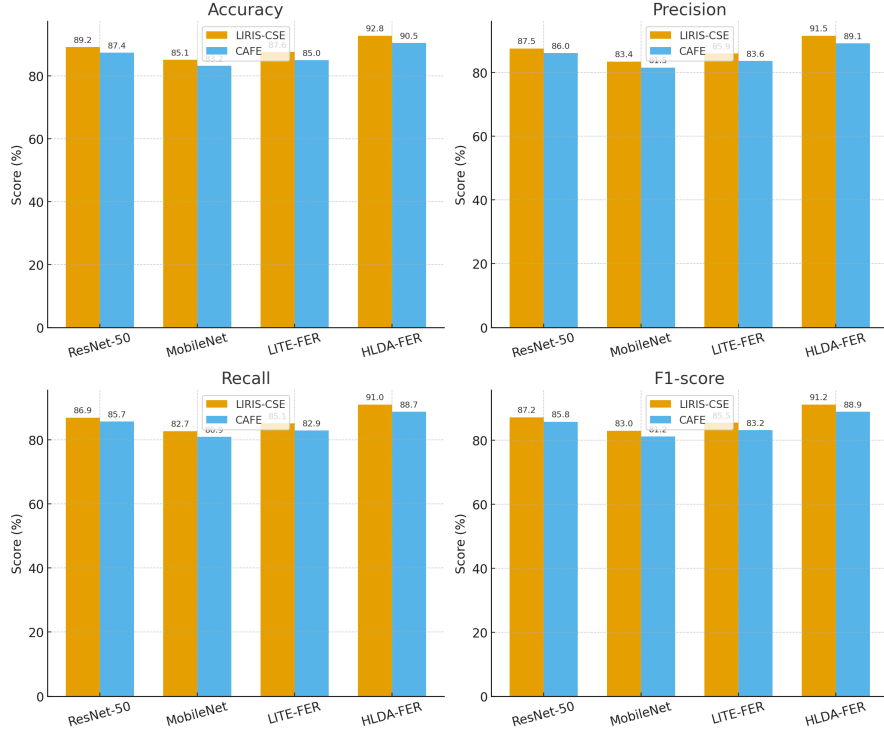


Fig. 3 Comparison of Accuracy, Precision, Recall, and F1-score for ResNet-50, MobileNet, LITE-FER, and HLDA-FER on LIRIS-CSE and CAFE datasets.

- [3] Laraib, U., Shaukat, A., Khan, R.A., Mustansar, Z., Akram, M.U., Asgher, U.: Recognition of children's facial expressions using deep learned features. *Electronics* **12**(11), 2416 (2023)
- [4] Vivekananthan, S.: Emotion classification of children expressions. arXiv preprint arXiv:2411.07708 (2024)
- [5] Bicer, E., Kose, H.: Lite-fer: A lightweight facial expression recognition framework for children in resource-limited devices. In: 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG), pp. 1–9 (2024). IEEE
- [6] Lee, E., Lee, E.-J., Anwar, S.M., Yoo, S.B.: Child fer: Domain-agnostic facial expression recognition in children using a secondary image diffusion model. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2750–2754 (2024). IEEE
- [7] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
- [8] Simonyan, K., Zisserman, A.: Very

- deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
- [9] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
- [10] Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* **34**, 8780–8794 (2021)
- [11] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. *Advances in neural information processing systems* **27** (2014)
- [12] Arora, N., Sharma, S.C.: A comparative study on facial image retrieval using local patterns. *Multimedia Tools and Applications* **83**(28), 70637–70692 (2024)
- [13] Arora, N., Sharma, S.C.: Etlbp and erdlbp descriptors for efficient facial image retrieval in cbir systems. *Multimedia Tools and Applications* **83**(4), 9817–9851 (2024)
- [14] Camras, L.A., Oster, H., Campos, J.J., Miyake, K., Bradshaw, D.: Japanese and american infants’ responses to arm restraint. *Developmental psychology* **28**(4), 578 (1992)
- [15] Liu, Y., Sourina, O., Nguyen, M.K.: Real-time eeg-based human emotion recognition and visualization. In: 2010 International Conference on Cyberworlds, pp. 262–269 (2010). IEEE
- [16] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
- [17] Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
- [18] LoBue, V., Thrasher, C.: The child affective facial expression (cafe) set: Validity and reliability from untrained adults. *Frontiers in psychology* **5**, 1532 (2015)