# Ensemble Models For Facial Emotion Recognition

Alexandru Berindeie

20 November 2024

**Abstract**

Facial Expression Recognition (FER) is a significant machine learning challenge, focused on the accurate identification of human facial expressions. Despite recent advancements, enhancing performance and obtaining reliable results in real-world conditions continues to pose substantial difficulties. This report presents several FER approaches, focusing on FER2013 dataset, a widely used dataset for this task. Firstly we present the problem that we are trying to solve then we present the theoretical background talking about related work, the notions used and about the dataset. We then present the methodology of research that we used to select the papers and the methods used. We present five popular and effective approaches focusing on their contribution and the results, some of those being ResEmoteNet, PAtt-Lite, Residual Masking Network. We implement them and try to improve the presented results using different augmentations on the dataset, changing hyper-parameters, employing an ensemble approach (composed of the best performing models) and even fine-tuning the models. Finally, we evaluate our method by comparing the results with leading models in the field, demonstrating a significant increase in accuracy. The results underscore the effectiveness of our approach, offering a comprehensive solution to the persistent challenges in FER, and paving the way for more empathetic AI, advanced human-computer interfaces, and improved mental health support systems.

**Keywords:** Facial Emotion Recognition, Deep-learning, Convolutional Neural Network, Residual Masking Network, Attention mechanisms, Squeeze and Excitation Network

## 1 Introduction

Facial Expression Recognition (FER) is a specialized domain within machine learning and computer vision that focuses on identifying and interpreting human facial expressions. By analyzing facial features such as the movement of muscles, changes in skin texture, and the spatial relationship between facial landmarks, FER systems aim to classify expressions into categories like happiness, sadness, anger, surprise, fear, and disgust. The process typically involves

several stages: face detection to locate faces within an image or video, feature extraction to identify key facial landmarks and regions, and classification using machine learning algorithms to determine the expression.

## 1.1 Motivation

Emotions are fundamental to human interaction, and deciphering them through facial cues has far-reaching implications. Accurate emotion recognition from facial expressions lays the groundwork for more empathetic AI, advanced human-computer interfaces, and improved mental health support systems. FER is a very important task in both computer vision and artificial intelligence, involving detecting and interpreting human emotions based on their facial expression. It has many applications, such as human-computer interaction, psychology, healthcare, and recommendation systems, and it's been a difficult task for a while because of the complexity of human facial expressions.

## 1.2 Research questions

The starting questions that guided this study are related to the way we can address some issues with FER datasets and how can we leverage the new mechanisms to increase models performance and efficiency.

Main research questions are:

- How can we address class imbalance and data limitations for FER datasets?
- What impact does ensembling of different FER models have on accuracy when compared to individual ones?
- How does the new approach compares with other leading models in the field?

To figure some of those questions, in this paper, we present and study several approaches that addresses two significant challenges in FER tasks: the class imbalance caused by a lack of negative emotion images and the limited size of the FER2013 dataset. We leverage recent advancements in generative artificial intelligence and FER tasks by using generative models to obtain the images needed to resolve the challenges presented above and by integrating several new and efficient FER models in an ensemble.

## 1.3 Structure of the report

The following sections are structured as follows: Section II present the problem that we try to solve along the challenges that appear. Section III begins with a presentation of the notions that will appear in this paper, then continues with the related work and ends with the presentation of the dataset that will be used. Next Section presents the methodology of research that was used, containing how we chose the papers, the approaches and the main criteria that we taken into account. Section V present the five approaches selected, starting with the

approach,the contribution, the architecture used and the results obtained by the authors. Section VI will present the implementations and the experiments that we did and the results that we obtained. Finally section VII will present the conclusion of our work along the limitations and the future work direction that we provide for future research in this area.

## 2 Problem formulation

The main **problem** is **facial emotion recognition based on images**. FER presents, as specified in Introduction, a significant challenge within the domains of machine learning and computer vision due to the complexity and variability of human emotions. One primary issue that makes this problem complex is the imbalance and limited size of datasets like FER2013, which hampers the performance and reliability of FER models.

An important aspect of the problem that we approach is accurately identifying human emotions from facial expressions with another aspect related to allowing it to be done in real-time so it can be useful in a wider area of applications. Additionally there is a secondary problem that appeared while trying to solve the first one, that is datasets imbalance issue that causes model to have a bias on the few classes that have more examples in training dataset and influences in a bad way its performance.

Additionally, FER tasks must contend with inter-class similarities and intra-class differences influenced by factors such as gender, age, and ethnicity, as well as occlusions, lighting variations, and changes in head pose. The challenges necessitate innovative solutions to improve the accuracy and robustness of FER systems in real-world conditions. Those (inter-class similarities and intra-class differences )also appear in [22], where they are discussed in more depth. Inter-class variations arise from differences in facial expressions among individuals, influenced by factors such as gender, age, and ethnicity. Conversely, intra-class variations include issues such as occlusions, varying illumination, and changes in head pose, all of which contribute to the complexity of accurately recognizing facial expressions.

Another part of the problem is the need for a system that can dynamically be adapted for various scenarios and contexts without increasing the hardware requirements too high so the solution can be versatile and useful for a wider variety of use cases.

An effective strategy for addressing some of these challenges, as stated in [25] and [6], involves concentrating on specific facial areas that provide crucial cues, such as the eyes, mouth, and eyebrows, while disregarding less relevant features like hair and the jawline. To this end, attention mechanisms have been introduced in the context of image classification [30], [33], aiming to enhance the performance of convolutional neural networks (CNNs) by enabling the model to focus on these critical details.

Another approach to achieving more robust and reliable results in FER is the use of ensemble models. This popular technique enhances the accuracy of FER

models by integrating multiple architectures, leading to more precise predictions. Ensemble models leverage the strengths of various algorithms, features, and datasets, often outperforming single models across diverse FER tasks. As technology advances, the ability to decode human emotions becomes increasingly sophisticated and accessible, paving the way for improved emotional understanding and enhanced quality of life.

# 3 Theoretical Background

## 3.1 Notions used

A masking network is a mechanism that focuses on specific regions of an input image [25] to enhance the network's performance. For FER tasks, masking often involves segmentation techniques to refine feature maps by highlighting critical facial regions and suppressing irrelevant areas. This approach helps the network focus on details like the eyes or mouth, which are crucial for expression analysis. Residual Masking Networks (RMNs) implement this through residual layers combined with masking blocks that learn attention scores, ensuring the preservation of valuable features while minimizing noise. Residual layers are a critical component of Residual Networks (ResNets), designed to address the problem of vanishing gradients in deep neural networks. They enable very deep networks to learn effectively by introducing skip connections, which bypass one or more layers in the network. Masking blocks are specialized mechanisms designed to improve the focus of a neural network on relevant features while ignoring irrelevant or noisy parts of the input. They are particularly useful in tasks like FER, where certain regions of an image contain more meaningful information than others.

Attention mechanisms [23] are computational strategies that allow a neural network to focus selectively on the most relevant parts of the input data, akin to how humans focus on specific details in a scene. These mechanisms have been widely used in FER to enhance feature learning. For instance, dot-product self-attention layers can capture long-range dependencies between features, while multi-head self-attention provides a diverse representation by attending to different parts of the input. Attention mechanisms are critical for processing complex images where fine-grained details influence the outcome. Self-attention layers are mechanisms that allow a neural network to focus on specific parts of an input sequence when computing its representation. They are commonly used in natural language processing and computer vision tasks, enabling models to capture relationships between elements in a sequence or spatial regions in an image. Multi-head self-attention extends self-attention by allowing the model to focus on multiple aspects of the input simultaneously. This is done by computing multiple independent self-attention operations (heads) and combining their outputs.

Squeeze-and-Excitation Networks [10] are architectures that enhance channel-wise feature representations by learning to recalibrate the importance of each

feature map. This involves a "squeeze" operation that aggregates global information and an "excitation" operation that generates adaptive weights for each channel. In FER, SE blocks are integrated to prioritize features representing key facial expressions, boosting the model's representational power while maintaining computational efficiency.

Residual Networks [8] are a deep learning architecture designed to address the vanishing gradient problem in very deep networks. ResNets introduce skip connections that allow the gradient to flow directly through layers, facilitating the learning of deeper representations. These skip connections also enable the network to reuse features, improving convergence and accuracy. For FER, ResNet-based blocks often serve as backbones in models to extract hierarchical features effectively.

Ensemble methods [3] combine predictions from multiple models to improve overall performance. In FER, ensemble techniques like simple averaging or weighted summation aggregate outputs from diverse architectures, boosting robustness and accuracy. Ensemble approaches help mitigate the limitations of individual networks and improve generalization.

Snapshot ensembles are a specific ensemble strategy where a single model is trained cyclically, saving its parameters at different points during the optimization process. These snapshots are treated as distinct models during inference, and their predictions are averaged for a final result. This method enhances the diversity of the ensemble without the need to train multiple models from scratch, making it a resource-efficient solution.

A Spatial Transformer Network [13] is a module that dynamically learns spatial transformations, such as scaling, rotation, or cropping, to align inputs to a canonical form. By focusing on key regions of an image, STNs improve feature extraction and robustness to spatial variations. The integration of STNs is generally done to handle challenges like facial rotations and varying expression scales effectively for FER tasks.

Patch extraction blocks [22] are specialized layers that divide input feature maps into smaller regions or patches for focused learning. These blocks are designed to capture localized high-level features. The patch extraction block splits input maps into smaller patches, processes them through depthwise separable convolutions, and applies global average pooling to reduce parameters and prevent overfitting. This lightweight yet effective approach contributes to the model's efficiency and accuracy.

Vision Transformers [31] are models that apply transformer architectures, originally developed for natural language processing, to image data. They split an image into patches, encode them as embeddings, and use multi-head attention to learn relationships between these patches. ViTs excel in capturing global context and have been adapted in frameworks which employs window-based cross-attention to enhance feature extraction.

## 3.2 Related work

In this chapter we will present some of the related work that we studied and showed very good results, find their strengths and element that we could apply in our approach.

In recent years, significant advancements have been made in the field of FER through the application of CNNs. Early works, such as LeNet-5 [15], laid the foundation for modern CNN architectures that have become pivotal in image classification tasks. Subsequent architectures like AlexNet [14], ResNet [8], and MobileNet [9] have demonstrated superior performance in FER by leveraging deep learning techniques. However, these models often require substantial computational resources, which may not be feasible for all applications.

To address this, a very good approach is PAtt-Lite [22] architecture that utilizes a truncated MobileNetV1 as its backbone, incorporating a lightweight patch extraction block and an attention classifier to enhance FER performance under challenging conditions. This method outperforms state-of-the-art approaches across multiple benchmark databases, including CK+ [19], RAF-DB [16], FER2013 [12], and FERPlus [24], with significantly fewer parameters. The key advantage of PAtt-Lite lies in its high accuracy despite its lightweight nature, making it a viable solution for resource-constrained environments.

The current state-of-the-art model, ResEmoteNet [27], introduces an innovative architecture that combines a CNN with three convolutional layers, a Squeeze and Excitation (SE) block, and three Residual (ResNet) blocks, resulting in a highly robust and efficient network. This design leverages CNN layers to progressively extract features, while the SE block enhances the network's ability to focus on critical facial features. The addition of Residual Networks enables the model to learn deeper, more complex data patterns. This architecture achieves outstanding results, setting new benchmarks on three widely used datasets: FER2013, RAF-DB, and AffectNet [21] with a substantial performance improvement but with a bigger dimension than PAtt-Lite.

Another notable development is the Residual Masking Network (RMN) proposed in [25], which integrates a novel masking mechanism to enhance the performance of CNNs in FER tasks. RMN employs a segmentation network to refine feature maps, focusing the network on crucial facial regions and thus improving classification accuracy [6]. By combining Deep Residual Networks with a U-net-like architecture [26], RMN achieves state-of-the-art results on the FER2013 and VEMO datasets, with notable accuracy improvements over models like ResAttNet56 and Densenet121 . These strengths, particularly the effective use of masking to enhance attention mechanisms and ensembling models (ensemble increased accuracy with 2.68% compared to single model), are elements we could apply in our approach to further improve FER performance).

Furthermore, the integration of handcrafted features with CNN architectures has shown promising results in FER [2]. Previous studies combined various CNN architectures with a single handcrafted model and employed a local learning strategy, leading to superior outcomes. This approach demonstrated significant improvements in [7], such as an 8% accuracy increase on the FER 2013 dataset.

The strength of combining deep learning and handcrafted features lies in leveraging the strengths of both methods, thus enhancing the overall recognition performance.

In addition to the previously presented approaches, the ARBEx framework [31] offers substantial improvements in Facial Expression Learning (FEL) through its innovative techniques. ARBEx utilizes a Vision Transformer (ViT) [4] with a window-based cross-attention mechanism to enhance feature extraction. It addresses issues like inter-class similarity, intra-class disparity, and label ambiguity through extensive data preprocessing, including heavy augmentation and data refinement. The use of trainable anchor points [20] and multi-head self-attention [32] for reliability balancing ensures robust label predictions and stabilizes class distributions.

Similarly, the EmoNeXt framework [5] adapts the ConvNeXt architecture [18] for FER by integrating a Spatial Transformer Network (STN) to focus on feature-rich regions to enhance the feature extraction efficiency and Squeeze-and-Excitation blocks [11] to capture channel-wise dependencies which help in recalibrating feature responses and improving the network's representational power. The introduction of a self-attention regularization term [29] further enhances the model's ability to generate compact feature vectors, resulting in superior performance on the FER2013 dataset compared to existing deep learning models. The combination of STN, SE blocks, and self-attention mechanisms in EmoNeXt highlights its comprehensive approach to enhancing both spatial and channel-wise feature representation. These elements collectively contribute to a more robust and efficient FER model, capable of delivering high accuracy in challenging conditions.

Facial area recognition is critical for improving FER systems. Traditional methods, such as the Haar Cascade classifier [28], have been widely used for face detection due to their simplicity and efficiency. The Haar Cascade classifier operates by scanning an image with a sliding window and applying a series of increasingly complex classifiers to detect the presence of facial features. These classifiers are trained using positive and negative samples, allowing the model to identify key facial structures such as the eyes, nose, and mouth. Despite its robustness in various lighting conditions and orientations, the Haar Cascade classifier can struggle with occlusions and variations in facial expressions.

More advanced techniques, such as Multi-task Cascaded Convolutional Networks (MTCNN) [17], improve upon traditional methods by leveraging deep learning. MTCNN uses a three-stage process to detect facial landmarks with high accuracy: a proposal network generates candidate windows, a refine network filters these windows to improve precision, and an output network produces the final bounding boxes and facial landmarks. This multi-stage approach provides effectively handling in challenging scenarios, including varying facial expressions, partial occlusions, and diverse poses.

### 3.3 Dataset

There are several datasets used for this task, out of those we decided to use FER2013 because of its size and popularity. The FER2013 dataset is a facial expression recognition dataset that contains 35,887 grayscale, 48x48 pixel images of faces, each labeled with one of seven emotion categories specified above. It is simple so it facilitates quick processing, makes efficient model training and is widely used as a benchmark in FER research providing a standard for comparison across studies.

For the dataset one of advantages that made us to decide to use it are its alignment with our approach's requirements, drawing from its widespread utilization within the research community. This dataset offers several other advantages crucial to our methodology, including its extensive usage across various studies, providing a diverse range of facial expressions essential for robust model training, its size that facilitates fast model training. However, it's worth noting that it also comes with certain limitations, such as its focus on facial expressions in controlled environments, the class imbalance (negative emotions number is much shorter than positive ones), the relatively modest size of the dataset when compared with larger ones such as AffectNet, FER+, and DISFA, and occasional labeling inaccuracies, characterized by either images that lack the facial expression completely or possess wrong labeling.

After we picked the dataset and studied its limitations, we decided to resolve two of the most important ones that are the class imbalance and the reduced size. To achieve that we needed to find how many images are needed to balance classes. For that purpose we created a python script (here) that analyze the dataset and generates a table detailing the distribution of images across different classes and calculating the number of images required for achieving equilibrium. The findings revealed the necessity to procure 21,796 images for training and 5,240 images for testing, increasing dataset size by 75%, initially it containing 35,887 images. This strategic augmentation not only rectified the class imbalance but also bolstered the dataset's size, thereby enhancing its utility for balanced and reliable models.

## 4 Methodology of Research

To address these challenges, our research adopts a comprehensive approach combining theoretical analysis, data augmentation, and empirical evaluation. The main steps are as follows:

Theoretical Analysis that consists of reviewing the existing literature on FER (Fig. 1), focusing on the strengths and weaknesses of various models and techniques. Then analyzing the specific challenges associated with FER datasets, such as class imbalance and dataset size limitations. A first step into our research is looking for sources of information where we can find papers on our subject and so the result of used resources is the following:

1. Google Scholar

2. arXiv
3. ResearchGate
4. SpringerLink
5. PapersWithCode
6. ScienceDirect

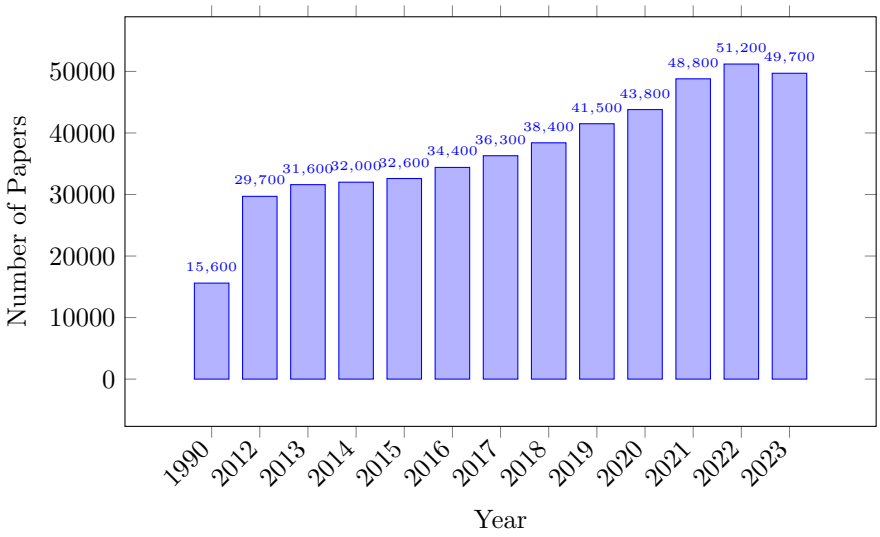



Figure 1: Number of Papers in Facial Emotion Recognition Over the Years

The next step was using keywords to find relevant paper to our domain and even filter the abundance of approaches and papers to find the ones that would help in our proposed approach. Some of the keywords used: facial emotion recognition, ensemble methods/models, deep learning, residual masking network, attention mechanisms, fer2013, raf-db.

The amount of results obtained was still to big so we included some inclusion/exclusion criteria such as publication date (as recent as possible), relevance (should be focused on the areas of research needed and results should be important), impact (should have several citations). We also taken into account the clarity of objectives and the problem addressed and the clarity of the solution.

A final filter of the selected papers was the comparison of their results with the results shown in PapersWithCode SOTA section for FER2013 dataset (here) and if results were at least close to those we took it into the consideration.

Dataset Analysis and Augmentation where we decided to use FER2013, identified its limitations, particularly the imbalance between positive and negative emotion classes. Then trough research we figured the solution of using text-to-image models to generate additional images, enhancing the dataset's size and

balancing it at the same time. In that process we developed a dynamic prompt generator to create diverse prompts that would help the model to generate realistic images that fit the FER2013 format and ensures consistent quality and variety.

Empirical Analysis consisted of training multiple FER models, including PAtt-Lite, Residual Masking Network and ResEmoteNet, on both the original and the augmented dataset. Combine these models into an ensemble to leverage their individual strengths and improve overall performance, providing at the same time the information to compare the baseline models with the newly trained ones.

Evaluation and Comparison included evaluating the models performance compared to original ones but also with state-of-the-art ones and analyzing the results to assess the effectiveness of the ensemble approach.

# 5 Papers studied

In this chapter we will present five FER approaches that we considered to be the most relevant for our purpose and that have the best results.

## 5.1 PAtt-Lite

**Problem:**
The addressed problem is FER, more specifically doing the task in the "wild" and under challenging conditions.
**Solution:**
The authors propose a lightweight patch and an attention network (Fig. 2) based on MobilNetV1 that is supposed to improve FER performance under challenging conditions. The model is built up upon a truncated MobileNetV1 combined with a patch extraction block and an attention classifier. Data goes firstly into MobileNetV1 to leverage the feature-extracting capabilities of the model, then the output is mapped from the patch extraction block. Attention classifier then takes the feature maps that were global average pooled and returns the probabilities of facial expressions. The patching block takes advantage of using a pre-trained CNN, authors cutting first 9 layers because first layers depthwise tend to learn generic features. Then the patch extraction block that is composed of three different convolutional layers, first two of them being depthwise separable and last one being a point-wise convolutional layer, operates on feature maps obtained from MobileNetV1 those being padded to dimension 16x16. First layer splits the feature maps in four patches while learns high-level features, then second layer and third layer learn higher-level features from patched feature maps resulting in a 2x2 output. Global average pooling is used to average the patch representation of patch extraction block so flattening is not necessary and so resulting in a reduction of parameters while also preventing over-fitting. Attention classifier is used for a better learning from the backbone of MobileNetV1 and the patch extraction block, it comprises a dot-product self-

attention layer placed between the two fully connected layers and the classifier.

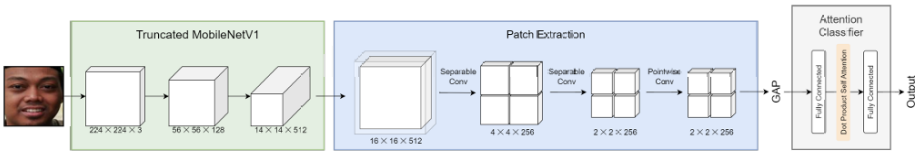



Figure 2: PAtt-Lite arhitecture.

**Results:**
The outcomes are showcased through the evaluation of the model on several datasets such as: CK+, Fer+, RAF-DB, Fer2013.

- **CK+:100% accuracy**
- **Fer+:95.55% accuracy**
- **RAF-DB:95.05% accuracy**
- **Fer2013:92.5% accuracy**

Another significant result (Table 1) of this approach is that obtained model has a relative low number of parameters compared with other approaches with worse performance.

| **Model** | **Number of parameters** |
|---|---|
| PAtt-Lite | 1.10M |
| VTFF | 80.1M |
| RAN | 11.2M |
| VTFF | 51.8M |
| SCAN-CCI | 70M |
| ARM | 11.2M |
| TransFER | 65.2M |
| Facial Chirality | 46.2M |
| APViT | 65.2M |
| POSTER | 71.8M |
| POSTER++ | 43.7M |
| CIAO | 17.9M |
| Imponderous Net | 1.45M |

Table 1: SoTA models and their number of parameters

**Open problems and future work:**
Through their research in the paper, the authors identified open issues and proposed future directions for their work.
Open problems:

- scarcity of negative expression images on the internet
- the class imbalance in the in-the-wild FER databases
- the lack of facial landmarks for FER
- the integration of other modalities such as audio and text for multi modal FER

Future work directions:

- using different backbone models
- using different patch sizes
- using different attention mechanisms
- using different loss functions

## 5.2 Residual Masking Network

**Problem:**
The authors address the problem of automatic FER and its applications in human-computer interaction and the challenges caused by variation in facial expression across individuals and situations. **Solution:**
The solution proposed makes use of a new masking idea that uses segmentation to refine the feature maps of a CNN so the model can focus on relevant information when trained. The model proposed (Fig. 3) by authors follows a Residual Masking Network (RMN) architecture that uses the masking idea and is applied as an ensemble method that combines different CNN models to improve overall performance. The flow presented starts with the residual masking blocks, their network contains 4 of those blocks each of them operating on different feature sizes and contains a masking block and a residual layer. The residual masking block performs the scoring operation, and uses a Resnet34 as the backbone. Residual layer takes care of feature processing and masking block takes care of producing weights. Masking block also uses attention residual learning that makes sure its not removing useful features and consists of a contracting patch and an expansive path. Ensemble method is also being used to furthermore increase accuracy, it takes Residual Masking Network and another 6 different CNN results as a simple non-weighted sum average as the final prediction.

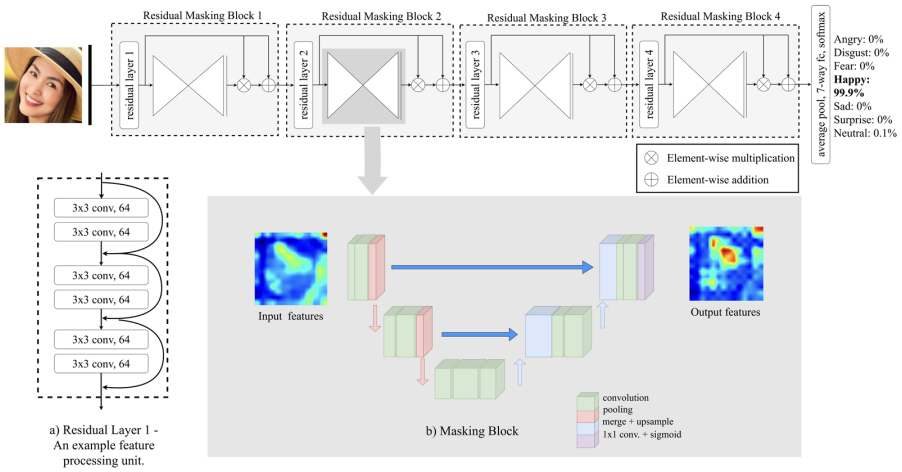



Figure 3: RMN arhitecture.

**Results:**
Obtaining model performance is done by evaluating it on two different datasets: Fer2013 and VEMO. For Fer2013 model performance is evaluated individually and using a simple weighted ensemble and obtains the following results:

- Without ensemble: 74.14 % accuracy
- Ensemble: 76.82 % accuracy

As for VEMO dataset, model obtains an accuracy of 65.94 % without ensemble, because authors considered there are not enough models with good results on that dataset to be worth using an ensemble.
**Open problems and future work:**
The open problems identified by authors are:

- The generalization ability on other tasks such as classification and detection.
- Parameter reduction and optimization to improve the network efficiency and speed.
- The evaluation on larger and more diverse datasets such as ImageNet.

Future work direction given by the authors:

- Testing the Residual Masking Network on the ImageNet dataset.
- To collect and annotate more facial expression images in the wild to create a larger and more balanced dataset for FER.

## 5.3 ResEmoteNet

**Problem:**
Facial emotion recognition is a challenging task due to the subtle changes in facial features and the complexities involved in capturing these details. Existing methods struggle with accurate recognition across diverse datasets, particularly when faced with issues such as mislabeling, inconsistent image quality, and varying conditions like pose and lighting.
**Solution:**
The authors propose ResEmoteNet, a novel deep learning architecture (Fig. 4) that integrates Convolutional Neural Networks (CNN), Squeeze-Excitation (SE) blocks, and Residual Networks to address these challenges. The SE blocks enhance feature representation by emphasizing important features while suppressing less relevant ones. The residual connections facilitate deeper learning by addressing gradient-related issues. The network also employs adaptive average pooling for consistent output sizes across diverse datasets. ResEmoteNet was trained with carefully tuned hyperparameters and data augmentation techniques to ensure robust performance.

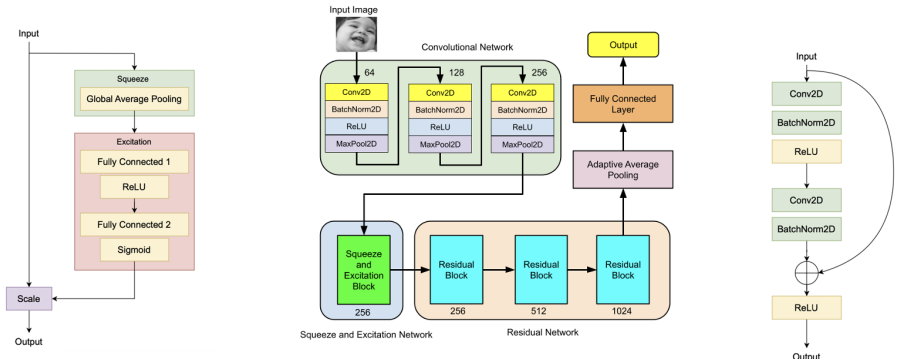



Figure 4: ResEmoteNet arhitecture.

**Results:**
ResEmoteNet was evaluated on three benchmark datasets—FER2013, RAF-DB, and AffectNet. It outperformed state-of-the-art models, achieving accuracy rates of 79.79%, 94.76%, and 72.93%, respectively. The inclusion of SE and residual blocks significantly improved feature extraction and classification performance. Confusion matrices for each dataset demonstrate the model's ability to reduce classification errors across various emotion classes.

**Open problems and future work:**
The open problems identified by authors are:

- Noisy and mislabeled datasets.
- Model size.
- Imbalance of the dataset.

Future work direction given by the authors:

- Enhancing robustness of the architecture.
- Testing on more datasets

## 5.4 ARBEx

**Problem:**
Facial Expression Recognition faces challenges including inter-class similarity, intra-class disparity, and data biases. Existing models, such as Vision Transformer (ViT)-based methods, fail to adequately address these issues, often leading to overfitting, label uncertainty, and poor predictions. These problems hinder effective recognition of subtle and complex emotional expressions, especially in diverse real-world conditions.
**Solution:**
The proposed ARBEx framework (Fig. 5) enhances FER by integrating a reliability balancing mechanism and advanced attentive feature extraction. ARBEx uses a Window-Based Cross-Attention ViT to extract multi-scale feature embeddings and employs trainable anchor points in the embedding space. Combined with a multi-head self-attention mechanism, ARBEx stabilizes label distributions and corrects unreliable predictions. Additional data preprocessing and augmentation techniques ensure robust performance. The model leverages anchor loss to separate embeddings, improving discriminative capabilities while maintaining stability in predictions.
**Results:**
Extensive evaluations demonstrate that ARBEx outperforms state-of-the-art models across multiple datasets, including RAF-DB, Aff-Wild2, and JAFFE. For example, ARBEx achieved:

- 72.48% accuracy on Aff-Wild2, surpassing POSTER++ by over 3%
- 92.47% accuracy on RAF-DB
- 96.67% accuracy on JAFFE

Having those results in mind we can see that it aligns with or exceeds other leading methods. ARBEx's reliability balancing significantly improved the stability and confidence of predictions. Visualization analyses, including plots, showcased better clustering and reduced inter-class confusion compared to competing models.

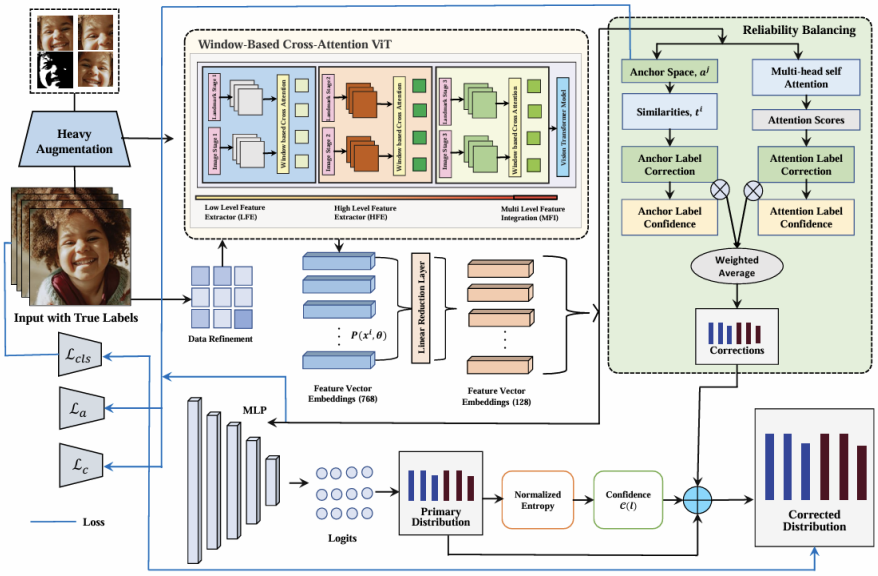



Figure 5: ARBEx arhitecture.

**Open problems and future work:**
The open problems identified by authors are:

- Computational complexity
- Mislabeled data in several datasets

Future work direction given by the authors:

- Extending ARBEx to handle multimodal tasks (2D and 3D)
- Refining label correction techniques for greater robustness against extreme noise and ambiguous labels
- Deploying ARBEx in real-world applications

## 5.5 EmoNeXt

**Problem:**
Facial Emotion Recognition is a challenging task due to variations in scale, rotation, lighting, and expression subtleties. Conventional methods often fail to capture these complexities, particularly when applied to unbalanced datasets like FER2013. Existing models lack the ability to dynamically adapt to spatial transformations or effectively emphasize relevant features, leading to suboptimal performance in emotion classification.

**Solution:**
The solution consists of the proposed EmoNeXt framework (Fig. 6) that enhances FER by adapting the ConvNeXt architecture with several innovative components:

- Spatial Transformer Networks (STN): These handle spatial variations by learning and applying transformations to input images, improving robustness to scale and rotation.
- Squeeze-and-Excitation (SE) Blocks: These recalibrate channel-wise feature maps, emphasizing critical information while suppressing noise.
- Self-Attention Regularization: This term balances the importance of features, ensuring compact and meaningful feature representations. The combination of these elements, coupled with advanced data augmentation and training techniques like AdamW optimizer and Mixed Precision Training, optimizes performance.

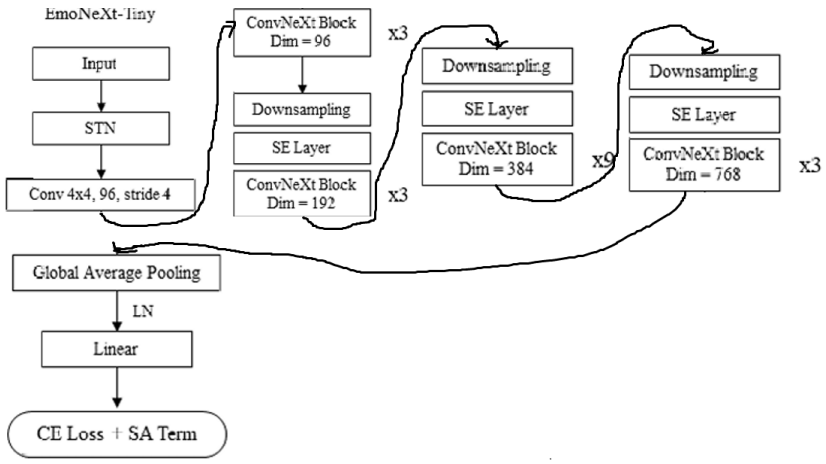



Figure 6: EmoNeXt arhitecture.

**Results:**
EmoNeXt demonstrated superior results on the FER2013 dataset, achieving a new state-of-the-art accuracy of 76.12% with its XLarge variant, outperforming previous models, including Segmentation VGG-19 (75.97%) and LHC-Net (74.42%). The Tiny variant of EmoNeXt also surpassed competitive models like ResNet50 and VGG with an accuracy of 73.34%, demonstrating its effectiveness across various model sizes.

**Open problems and future work:**
The open problems identified by authors are:

- Exploring integration with multimodal data (voice)
- Emotion recognition using different types of data

Future work direction given by the authors:

- Enhancing model generalization to other FER datasets with different distributions
- Refining label correction techniques for greater robustness against extreme noise and ambiguous labels
- Extending applications to real-world scenarios

# 6 Implementation and experiments

## 6.1 Proposed Approach

Our proposed approach comprises two significant components, whose integration yields the ultimate outcome, an ensemble that incorporates innovative methods and is trained on a balanced dataset. The code for data augmentation and data processing is provided here.

The first component, illustrated in Fig. 7, addresses the crucial task of dataset balancing, essential for training the models.

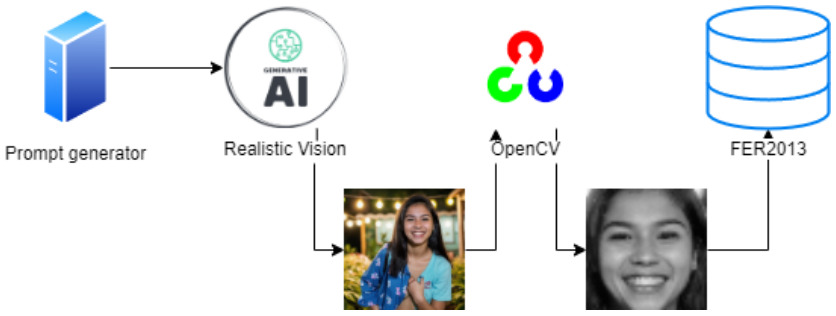



Figure 7: Dataset balancing

The second component, depicted in Fig. 8, presents the process of deriving the final prediction through an ensemble of models, each trained on the balanced dataset, aggregating their individual predictions.

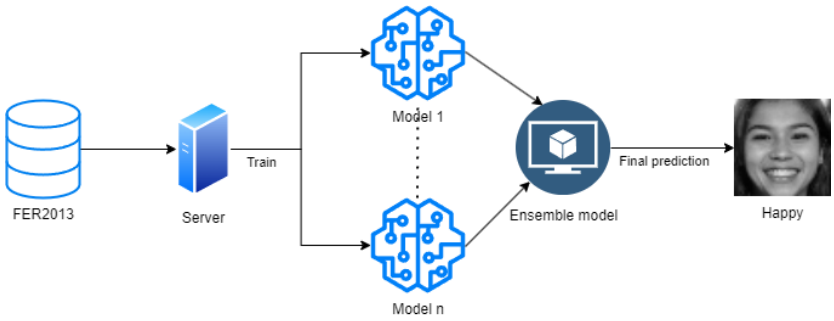



Figure 8: Creating final model

## 6.2 Image generation

Now that we've determined the number of images we need, the next step is to generate them efficiently. To do this, we must choose a model that produces realistic images quickly. We tested several popular models, including Dall-E, Stable Diffusion v1.4, v1.5, and XL, as well as Realistic Vision v4 and v6_b1.

After some trial and error, we discovered that the best results came from generating images with a minimum resolution of 512x512 pixels. Some models required 1024x1024 pixels for optimal quality, but this significantly increased the generation time. By adjusting the generation parameters, we achieved stable and satisfactory results with a resolution of 512x512 pixels, 40 inference steps, a guidance scale of 5.5, and specific negative prompts to stabilize the final output.

We ultimately selected Realistic Vision v6 because it offered the most advantages and met our needs perfectly. It is relatively small in size, focuses on generating realistic images, is easy to configure using the Hugging Face pipeline, and produces good results at a reduced image size. Other models had drawbacks: while Dall-E produced excellent results, it is only available as a paid version (the public version has been removed). Stable Diffusion XL had a large size and didn't perform well at 512x512 resolution, and Stable Diffusion v1.5 was outperformed by Realistic Vision v6 for our specific requirements.

We encountered another problem: using a simple prompt and adjusting the model's seed wasn't enough to generate a satisfactory variety of images. The model tended to produce images of the same person in different contexts or with only minor changes. To overcome this, we implemented a dynamic prompt generator. This tool mixes various elements of the main prompt and allows users to choose the desired emotion, number of samples, and starting seed value.

The main prompt format is as follows: [selected-context] image with a [emotion1], [emotion2], deeply [emotion3] single [selected-gender] that is [selected-emotion] and has a detailed [selected-expression]. Person is expressive, [selected-age], [selected-nationality] and has a [selected-race] genetic. Focus on eyes, eyebrows, mouth and other facial factors to express the [emotion1]. Each element is randomly chosen from a pool of values for each generated image, and the seed

is increased by 200 after each image to ensure greater variety.

Next, we needed to format the new images to be compatible with the FER2013 dataset and address issues such as non-close-up images, background elements, body parts, and some NSFW content that resulted in blank black images. We developed a preprocessing function to handle these issues:

1. Load a face detector from the OpenCV library.
2. Load the image and convert it to grayscale.
3. Use the face detector to extract only the face from the image.
4. Resize the extracted face to 48x48 pixels and save it to the output directory.

This process ensured that the newly generated images contained only the face and were in the desired format to fit in the original dataset.

We also created a Python script to count the images in each category and class in the FER2013 dataset, calculate how many images were needed to balance each class, and copy images from the generated directories to the original dataset. The script then verified that all classes were balanced by displaying the image count for each class.

Finally, we developed a Python script to save the dataset in a simpler, more accessible format that loads faster for model training. Using the h5py library, we created an .h5 file containing the resized images, along with their emotion indices, split into training, testing, and validation categories. This step was necessary because some models we plan to train require different image sizes than those in the original FER2013 dataset and need fast loading times due to the extended training process.

## 6.3 Implementation

In this section we'll present the starting points of our approach and initial results for its two main parts: the dataset and the model along some details like the setup, visualization of training process and some comparisons with baseline models. The part on which we **focus**, that is also the **target** of this **report** is second part related to the model where we present the implementation and the experiments.

### 6.3.1 Dataset

For the dataset part, we start with the results obtained by several models that use prompts to generate images, results that can be seen bellow in Table 2. We show the comparison result of the best two models used, that were configured to generate images based on the same prompt and same parameters. We compare the quality of image generated and the time taken to generate it, experiment is done on local machine using a GeForce RTX 4090 (laptop version) Graphics Cards and 32GB DDR5-4800 RAM. As we can see in table Realistic

Vision Model have clearly more human accurate images and the generation time is 5x times faster, the reason for speed is that Stable Diffusion XL generates 1024x1024 pixels images for the best results when Realistic Vision obtains good results even with 512x512. For the time we generated an batch of 5 images based on same promt and repeated test for 3 times then we averaged the times of each model.

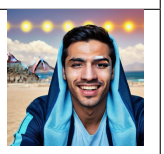
| Model | Generated Images | | |
|---|---|---|---|
| **Realistic Vision** 2.13 sec./img. |  |  |  |
| **Stable Diffusion XL** 11.37 sec./img. |  |  |  |

Table 2: Models performance and average image generation time

We also tested with other models, even paid ones like Dall-E but since those are not working on local machine the process to get the image on local took more time (model generates the image trough an API and then downloads it to local device) and of course it was more expensive, final result not being better than with the model we picked.

### 6.3.2 Model

For the model part we start the process by testing several models from whom we chose the best three for our purpose based on their results, their advantages, our needs and those are [27], [25] and [22]. First step in this direction was to train the models proposed on the original FER2013 dataset, to test their performance and to have a comparison data for later.

As a starting point of this work to highlight the obtained results utility, the human capacity of accurately recognizing the emotions of the people in the images of the dataset correctly is around 65 $\pm$ 5% as stated in [1].

The **implementation** and **experiments** on the latest state-of-the-art model on FER2013 dataset are available at **implementation**.

For ResEmoteNet model initial results been close to those presented by the authors when we checked with their checkpoint (78.08% test accuracy and 78.23% on validation accuracy). When we tried to train our own model we did not manage to pass 64% accuracy on either test or validation sets so we tried several experiments to improve the performance.

First experiment consisted on switching optimization function (we tried AdamW, RAdam, RangerAdaBelief) and adding a scheduler for adaptation of learning rate (we tried CyclicLR, CosineAnnealingWarmRestarts) with the

hope of improving accuracy and generalization capabilities of the model. The accuracy improved but still wasn't close to the one specified by the authors.

Then we employed an snapshot ensemble with a cyclic learning rate based on checkpoints of the model at different epochs during training (30, 40, 50, 60, 70, 80), the one with best test accuracy and with the best validation accuracy to see if we manage to improve the performance further. We managed to increase performance but we concluded that the size of the ensemble was not worth the accuracy improvement (around 2% accuracy for the 5 * size) so we moved to the final experiment that is fine-tuning the model.

Finally we fine-tuned the model on train part of the dataset, then on a mix of train and validation and then only on validation. We added here more data augmentation techniques, tried with different optimizers and schedulers but the accuracy still did not get close to the one claimed by the authors, we got around 65% at best. After some investigations we concluded that the authors used extra training data that is not public or clearly specified anywhere.

Initial PaTT-Lite model results show that it achieves an accuracy of 77% on the original dataset, that's considerably lower than authors present in their work (they claim it achieves 92.5%). After some research on their approach and their GitHub project page, the reason for that difference is the hyperparameters that authors used, the dataset augmentation(they did some changes on the original one, like image filtering) and used additional training data.

For the ensemble proposed in RMN initial results were more complex since the authors used several models for their approach, so we chose top 3 ones based on accuracy. For those models we trained them on the original dataset and obtained the following results:

1. RMN: 74.14%
2. Cbam_Resnet:73.34%
3. ResNet:73.14%

# 7 Final results and comparisons

Text for results and comparisons.

# 8 Discussion

## 8.1 Conclusion

Conclusion text.

## 8.2 Limitations and future work directions

Future work directions text.

# References


[1] Christian Białek, Andrzej Matiolański, and Michał Grega. An efficient approach to face emotion recognition with convolutional neural networks. *Electronics*, 12(12):2707, 2023.

[2] Tee Connie, Mundher Al-Shabi, Wooi Ping Cheah, and Michael Goh. Facial expression recognition using a hybrid cnn–sift aggregator. In *International workshop on multi-disciplinary trends in artificial intelligence*, pages 139–149. Springer, 2017.

[3] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Yassine El Boudouri and Amine Bohi. Emonext: an adapted convnext for facial emotion recognition. In *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2023.

[6] Yingruo Fan, Jacqueline CK Lam, and Victor OK Li. Multi-region ensemble convolutional neural network for facial expression recognition. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I 27*, pages 84–94. Springer, 2018.

[7] Mariana-Iuliana Georgescu, Radu Tudor Ionescu, and Marius Popescu. Local learning with deep and handcrafted features for facial expression recognition. *IEEE Access*, 7:64827–64836, 2019.

[8] 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*, pages 770–778, 2016.

[9] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[12] Yoshua Bengio Ian Goodfellow, Will Cukierski. Challenges in representation learning: Facial expression recognition challenge, 2013. Data set.

[13] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[15] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[16] Shan Li, Weihong Deng, and JunPing 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*, pages 2852–2861, 2017.

[17] Xiaochao Li, Zhenjie Yang, and Hongwei Wu. Face detection based on receptive field enhanced multi-task cascaded convolutional neural networks. *IEEE Access*, 8:174922–174930, 2020.

[18] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[19] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010.

[20] Jiawei Mao, Rui Xu, Xuesong Yin, Yuanqi Chang, Binling Nie, and Aibin Huang. Poster++: A simpler and stronger facial expression recognition network. *arXiv preprint arXiv:2301.12149*, 2023.

[21] 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*, 10(1):18–31, 2017.

[22] Jia Le Ngwe, Kian Ming Lim, Chin Poo Lee, and Thian Song Ong. Patt-lite: Lightweight patch and attention mobilenet for challenging facial expression recognition. *arXiv preprint arXiv:2306.09626*, 2023.

[23] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.

[24] Microsoft + 5 other contributors. Challenges in representation learning: Facial expression recognition challenge, 2016. Data set.

[25] Luan Pham, The Huynh Vu, and Tuan Anh Tran. Facial expression recognition using residual masking network. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4513–4519, 2021.

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[27] Arnab Kumar Roy, Hemant Kumar Kathania, Adhitiya Sharma, Abhishek Dey, and Md Sarfaraj Alam Ansari. Resemotenet: Bridging accuracy and loss reduction in facial emotion recognition. *arXiv preprint arXiv:2409.10545*, 2024.

[28] Sander Soo. Object detection using haar-cascade classifier. *Institute of Computer Science, University of Tartu*, 2(3):1–12, 2014.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[30] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017.

[31] Azmine Toushik Wasi, Karlo Šerbetar, Raima Islam, Taki Hasan Rafi, and Dong-Kyu Chae. Arbex: Attentive feature extraction with reliability balancing for robust facial expression learning. *arXiv preprint arXiv:2305.01486*, 2023.

[32] Zhengyao Wen, Wenzhong Lin, Tao Wang, and Ge Xu. Distract your attention: Multi-head cross attention network for facial expression recognition. *Biomimetics*, 8(2):199, 2023.

[33] 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)*, pages 3–19, 2018.