Facial Emotion Recognition

Jerome Agius
University of Malta
jerome.agius.21@um.edu.mt

Introduction

Facial Emotion Recognition (FER) is a specialised subfield of computer vision focused on automatically detecting human emotions from facial images or videos. Unlike generic image recognition tasks, FER demands precise interpretation of dynamic and subtle facial movements, such as eyebrow raises, lip curvature, or eye widening, which collectively convey emotions like happiness, anger, or surprise. This task is inherently complex due to the inter subject variability in facial expressions (e.g., cultural differences, individual idiosyncrasies) and environmental factors (e.g., lighting variations, occlusions). Moreover, practical implementations face challenges such as biased or limited training datasets, ethical concerns around privacy, and the difficulty of generalising models to real world, unconstrained scenarios.

Despite these hurdles, FER holds significant societal value. It enables advancements in human computer interaction (e.g., empathetic AI assistants), mental health diagnostics (e.g., depression detection), and public safety (e.g., identify emotions triggering potential terrorism threats) [1]. This paper conducts a systematic survey of state-of-the-art (SOTA) deep learning techniques for FER, outlining their techniques, strengths and limitations whilst deriving a conclusion on the best overall model.

Literature Review

Facial Emotion Recognition (FER) systems typically follow a three stage pipeline comprising face detection, facial expression detection, and emotion classification. Although this pipeline underpins the majority of FER implementations, numerous architectures have been proposed to overcome inherent challenges, such as variations in illumination, facial pose, and occlusion. Furthermore, FER methods can be broadly classified based on the nature of input data; static or dynamic corresponding to images and videos, respectively with the former serving as the primary focus for this research.

A comprehensive review by Rehman et al. [2] offers a broad analysis of FER methodologies, encompassing traditional machine learning, deep learning, and hybrid approaches. Rather than conducting original experiments, the authors compiled and synthesised findings from existing research to draw comparisons between commonly used models. Within the scope of deep learning, they discussed prominent architectures such as MobileNet, Convolutional Neural Networks (CNNs), Deep Convolutional Neural Networks (DCNNs), VGG16, and ResNet-50. The review examined how these models performed across several benchmark datasets, including CK+, JAFFE, FERPlus, FER2013, KDEF, RAF-DB, MUG, MMI, and AffectNet, highlighting key architectural choices and reported outcomes.

Accuracy emerged as the most frequently used metric for evaluating model performance across the referenced studies. Although additional metrics such as F1-score, precision, and recall were mentioned, their application was limited to specific contexts, revealing a lack of consistency in evaluation criteria. This inconsistency points to the need for more standardised and comprehensive benchmarking practices in FER research.

Based on the findings derived in the review, deep learning models generally outperformed traditional machine learning methods such as Support Vector Machines (SVMs) and Random Forests (RFs) especially in recognising subtle emotional expressions across diverse datasets. The review also noted that hybrid methods, which integrate traditional feature engineering with end to end deep learning, offer potential advantages in addressing issues like limited data availability and poor generalisation. Additionally, the authors emphasised the importance of data augmentation techniques including rotation, scaling, colour variation, and other image transformations for enhancing model robustness. These strategies help to increase the diversity of training samples, thereby reducing over fitting and improving general performance.

In a similar vein Suresh, Yeo, and Ong in [3] conducted a comparative study of three models ResNet50, Inception-ResNet, and ResNet50 with Entropy Regularisation to determine the most effective model for FER, with a focus on generalisability across twelve datasets. These included six in-lab datasets (JAFFE, CK+, Oulu-CASIA, KDEF, IASLab, GEMEP) and six in-the-wild datasets (EmotioNet, SFEW, RAF-DB, Aff-Wild2, FER2013, AffectNet). The comparison was conducted through two main experiments, using accuracy as the sole performance metric to evaluate both single source and multi source generalisation.

Before the experiments, all datasets were retrieved and preprocessed. This involved facial detection and alignment using the MTCNN model, greyscaling, image resizing, and data augmentation through a 50% horizontal flip. To ensure consistent comparison, all models were trained using identical hyperparameters across experiments.

In the first experiment, each model was trained on a single dataset at a time and then tested on all remaining datasets. This setup aimed to assess the generalisability of each model across different datasets. In the second experiment, models were trained on all but one dataset within the same context (either in-lab or in-the-wild), and then tested on the excluded dataset as well as on datasets from the opposite context. In the final phase of this experiment, each model was trained on eleven of the twelve datasets, with the remaining one used as a test set following a leave one out approach.

The results revealed that average single source within-corpus performance was approximately 76.4%, which dropped significantly to 42% when evaluating cross-corpus generalisation. This highlighted the difficulty of generalising from one dataset to another. In the multi source setting, within-setting generalisation yielded an average accuracy of 61.7%, while cross-setting performance was lower at 42.76%, showing that training on multiple same setting datasets offers a clear benefit over single source training, though it still struggles to generalise across different domains.

Finally, the leave-one-out setup achieved an average accuracy of around 65.6%, suggesting that the best generalisability is achieved when models are trained on a diverse mix of datasets spanning multiple domains. In addition to identifying optimal training strategies for generalisation, the authors noted that models trained on in-the-wild datasets consistently performed well across different test sets. This was attributed to the typically larger size and variability of in-the-wild datasets compared to their in-lab counterparts.

Following the structure of Rehman et al.'s review, the subsequent subsections detail the datasets and model architectures relevant to this research paper.

A. FER Benchmark Datasets

FER2013: Curated for the 2013 Kaggle competition "Challenges in Representation Learning: Facial Expression Recognition Challenge." FER2013 contains approximately 35,000 48×48 greyscale images labelled with seven primary emotions: Anger, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise. Despite its widespread use, the dataset suffers from class imbalance; for instance, the Disgust class has only 547 samples, while the Neutral class has 6,198 samples [4].

RAF-DB: The Real-world Affective Faces Database (RAF-DB) comprises around 30,000 internet sourced facial images, each annotated by 40 individuals to ensure label reliability. Images are standardised to 100×100 pixels and exhibit variation in age, gender, ethnicity, head pose, lighting, occlusion, and post-processing. Two annotation schemes are provided: basic emotion labels (Anger, Disgust, Fear, Happiness, Sadness, and Surprise) and compound expressions (e.g., Fearfully Surprised, Sadly Angry), making it well suited for both fundamental and nuanced emotion recognition tasks [5].

AffectNet: AffectNet is one of the largest available FER datasets, containing over 1 million facial images collected via multilingual keyword searches across three search engines. Approximately 440,000 images were manually annotated into eight categories: Neutral, Happy, Sad, Surprise, Fear, Anger, Disgust, and Contempt. Each image is also assigned valence and arousal scores generated via deep neural networks, supporting both categorical and dimensional emotion analysis [6].

CK+: The Extended Cohn-Kanade (CK+) dataset comprises 593 video sequences from 123 subjects, each beginning with a neutral expression and culminating in one of seven peak expressions: Anger, Contempt, Disgust, Fear, Happiness, Sadness, and Surprise. While primarily a video dataset, static FER studies often extract the peak expression frames, thus deriving a static dataset. Additionally, its controlled settings (consistent lighting, frontal poses, minimal occlusion) make it ideal for evaluating FER models under ideal conditions [7].

JAFFE: The Japanese Female Facial Expression (JAFFE) dataset includes 213 greyscale images (256×256 pixels) of posed expressions from 10 Japanese women. Each image is annotated with one of six basic emotions Anger, Disgust, Fear, Happiness, Sadness, and Surprise and rated semantically from 1 (low) to 5 (high) across annotators. Its controlled acquisition makes it useful for studies examining gender or cultural aspects of emotion expression [8, 9].

B. FER Benchmark Models

EmoNeXt: EmoNeXt builds upon Facebook's ConvNeXt backbone [10], introducing three key enhancements to improve facial emotion recognition. These include a Spatial Transformer Network (STN), Squeeze-and-Excitation (SE) blocks, and a self regularisation mechanism integrated into the loss function. The complete architecture is illustrated in Figure 3.

The STN module enhances image alignment by correcting distortions such as scale, orientation, and positional variance before further processing. It comprises a localisation network that predicts transformation parameters, a grid generator that constructs a sampling grid, and a sampler that applies the predicted transformation to the input image. This process improves the model's focus on relevant facial regions and is detailed in Figure 1.

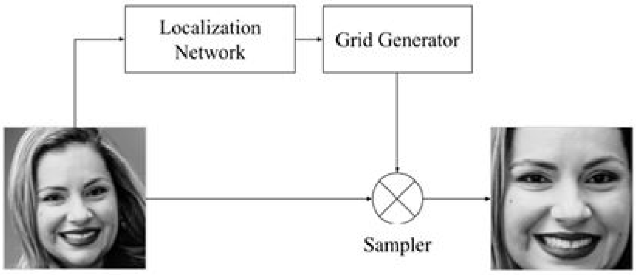

Fig. 1. Spatial Transform Module architecture [10]

SE blocks, applied after downsampling, enable the model to emphasise critical features by modelling channel wise dependencies. This is achieved through a three step process: global average pooling compresses each feature map (squeeze), a small fully connected network estimates channel importance (excitation), and the learned weights are used to rescale the original feature maps (scale). This mechanism is depicted in Figure 2.

Finally, the self attention regularisation term promotes balanced feature representation by minimising the variance of attention weights from their mean. The self attention mechanism

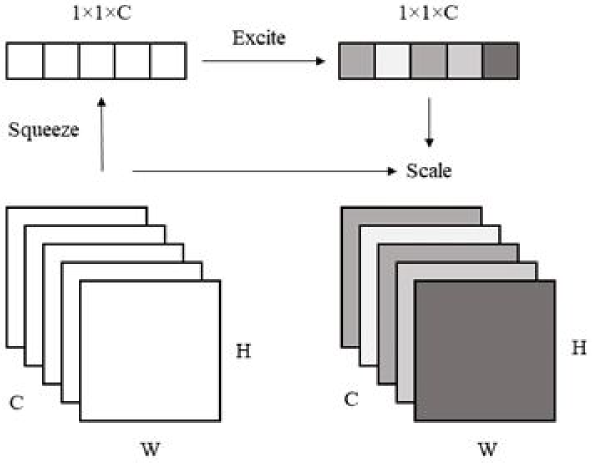

Fig. 2. Squeeze-and-Excitation process [10]

first computes pairwise similarities between queries and keys using a scaled dot product and softmax. The regularisation term then measures the deviation of each attention weight from the mean, encouraging compactness in the feature space. The total loss function is defined as $L_{final} = L_{CE} + \lambda \cdot L_{SA}$, combining the standard cross entropy loss with the self attention regularisation term to enhance generalisation.

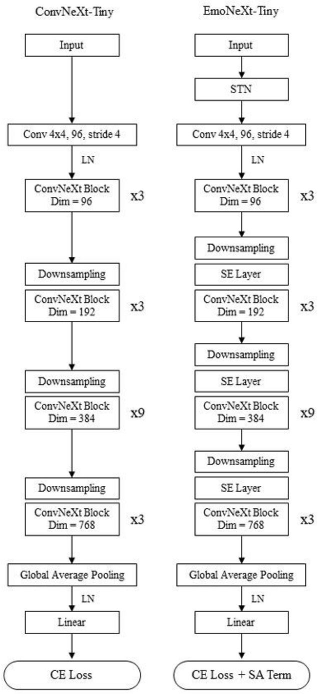

Fig. 3. ConvNeXt & EmoNeXt Architecture Designs

ResEmoteNet: The ResEmoteNet architecture as outliend in [11] integrates Convolutional Neural Networks (CNNs), Residual connections, and Squeeze-and-Excitation (SE) blocks to robustly capture facial emotion features. Unlike EmoNeXt, ResEmoteNet builds on a traditional CNN backbone enhanced with residual blocks and channel wise attention, achieving state-of-the-art performance across four open source emotion recognition datasets:

- **FER2013**: 79.79% accuracy
- **RAF-DB**: 94.76% accuracy
- **AffectNet-7**: 72.39% accuracy
- **ExpW (Expressions in the Wild)**: 75.67% accuracy

The model is composed of a CNN comprising three convolutional layers for hierarchical feature extraction where each layer is followed by batch normalisation and max pooling aimed at stabilising training whilst reducing spatial dimensions and enhancing robustness. The resultant feature maps are then fed into the SE block, which follows the standard Squeeze-and-Excitation mechanism (global average pooling + two FC layers with ReLU/sigmoid). The last primary component is a Residual Network consisting of three residual blocks with skip connections aimed at mitigating vanishing gradients and degradation in deep networks. The result from the residual network is then processed via Adaptive Average Pooling (AAP), which dynamically adjusts kernel size and stride to produce a fixed size output regardless of input dimensions, unlike traditional pooling methods that reduce spatial dimensions. The pooled features are then passed to the classifier for emotion prediction. The model architecture, including its CNN, SE, and residual components, is illustrated in Figure 4.

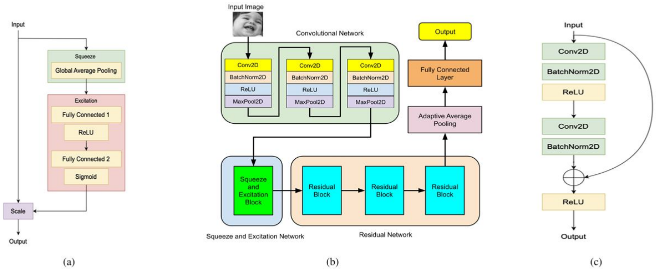

Fig. 4. ResEmoteNet Architecture [11]

ResNet50: The ResNet50 architecture, first introduced in [12], serves as a foundational deep learning model that has significantly influenced modern architectures. ResNet (Residual Networks) addresses the vanishing gradient problem common in deep networks by incorporating residual learning (skip connections that bypass one or more layers) thereby enabling efficient training even with hundreds or thousands of layers.

ResNet architectures consist of several variants, including ResNet18, ResNet26, ResNet34, ResNet50, ResNet101, and ResNet152. Although these models share the same fundamental design, they differ in depth and complexity, with ResNet50 being one of the most popular due to its balance between accuracy and computational efficiency.

ResNet50 is a 50 layer deep convolutional neural network composed of five main stages. The network begins with a 7×7 convolutional layer with a stride of 2, followed by a max pooling layer to reduce spatial dimensions early on. This is followed up with several sequential residual blocks which facilitates efficient feature extraction while keeping computational costs low, with each block consisting of three layers:

- A 1×1 convolution for dimensionality reduction.
- A 3×3 convolution for spatial feature extraction.
- A 1×1 convolution for restoring channel dimensions.

In subsequent stages, downsampling is performed directly by convolutional layers with a stride of 2, rather than dedicated pooling layers. This approach allows the network to effectively reduce the spatial resolution while doubling the number of filters when necessary, preserving computational efficiency. Finally, a global average pooling layer aggregates the spatial information before the fully connected layer produces the final class predictions.

This design reduces computational complexity while maintaining high representational power, making ResNet particularly well suited for large scale image recognition tasks. The architecture is illustrated in Figure 5.

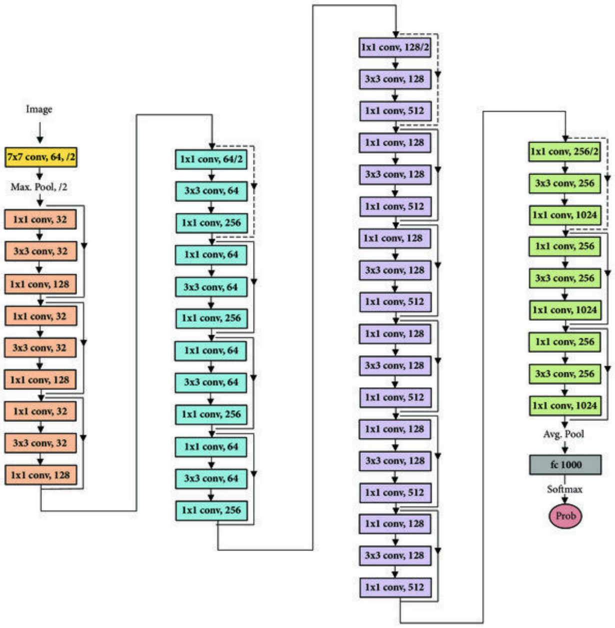

Fig. 5. ResNet50 Architecture [13]

DDAMFN: The Dual-Direction Attention Mixed Feature Network (DDAMFN) is composed of two components the Mixed Feature Network (MFN), itself adapted from MobileFaceNet and a the Dual-Direction Attention Network. By leveraging attention mechanisms, DDAMFN tackles both small inter-class differences and large intra-class variations through enhanced connections among different facial regions.

The MFN produces foundational feature maps from the input images and is composed of two principal block types:

- **Residual Bottleneck Block:** Uses shortcut connections to capture complex features and improve gradient flow, mitigating the degradation problem in deep networks.
- **Non-Residual Block:** Increases the network's ability to learn diverse and discriminative facial features.

To further improve performance, MFN integrates multiple strategies such as:

- **MixConv Operation:** Utilises multiple kernel sizes simultaneously within each bottleneck to capture a broader range of features.
- **Activation and Depth Adjustments:** The PReLU activation function is chosen over ReLU as it allows enhanced feature extraction whilst the network depth is tuned to prevent overfitting.
- **Coordinate Attention:** Refines feature extraction by focusing on important spatial locations in the feature maps.

Building on these feature maps, the Dual-Direction Attention Network (DDAN) applies multiple attention heads that generate horizontal and vertical direction feature maps. Instead of average pooling, a linear GDConv layer is employed to assign distinct importance to different spatial positions. The attention maps are combined element wise to form a final attention map, and the most informative map among the multiple heads is selected to highlight critical facial regions. An additional attention loss (based on Mean Squared Error between attention maps from different heads) ensures each head targets distinct facial areas.

Following the DDAN, the resulting 7×7×512 feature map passes through a linear GDConv layer and is then reshaped into a 512 dimensional vector. A fully connected layer produces the final class predictions. The model is trained with a composite loss function, combining standard cross entropy loss (L_{cls}) and an attention loss ($Latt$), where the overall loss is defined as: $L = L_{\text{cls}} + \alpha L_{\text{att}}$ with α typically set to 0.1. By integrating MFN and DDAN, DDAMFN effectively balances model complexity and generalisation, yielding improved performance on FER tasks. The model architecture is illustrated in Figure 6.

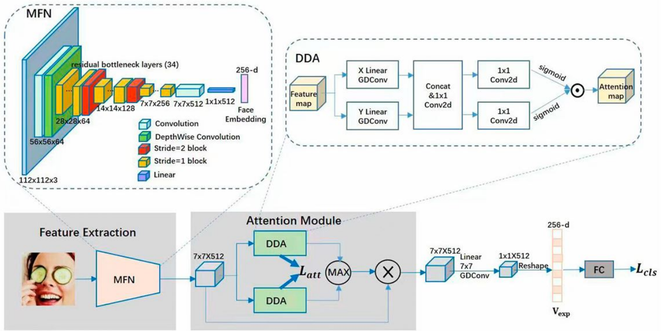

Fig. 6. DDAMFN Architecture [14]

PAtt-Lite: The PAtt-Lite architecture, introduced in [15], consists of three main components. The first component is the truncated MobileNetV1 model, which includes all layers up to the depthwise convolution in block nine. This truncation enables the model to leverage MobileNetV1's feature extraction capabilities while maintaining a lightweight structure. The second component is the patch extraction block, consisting of three convolutional layers: the first two are depthwise separable convolutions, while the final layer is a pointwise convolution. The first separable convolution layer splits the feature maps into four patches, learning higher level features from the input, while the subsequent layers refine these features further. This design enhances classification performance on challenging subsets and reduces the number of model parameters.

Next, Global Average Pooling (GAP) is applied to address overfitting. GAP computes a single average value for each feature map, resulting in a smaller output volume. This, in turn, reduces the number of parameters and minimises the risk of overfitting. Finally, the attention classifier, consisting of a dot product self attention layer placed between two fully connected layers, enables the network to learn attention weights. These weights highlight the most relevant parts of the input, thereby improving classification performance. The architecture described above is depicted in Figure 7.

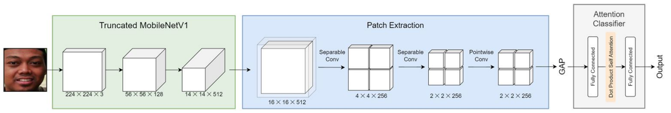

Fig. 7. PAtt-Lite Architecture [15]

C. FER Comparative Metrics

Among the various studies comparing FER models [2, 3], one primary evaluation metric stands out: accuracy scarcly accompanied by precision, recall, and F1-score, alongside a confusion matrix for deeper analysis.

Beyond these evaluation metrics, generalisability is a crucial factor in assessing FER models, as they must perform well on unseen data to be applicable in real world scenarios [3]. Generalisability can be categorised into single source and multi source generalisability, which denote whether a model is trained on a single dataset or multiple datasets, respectively.

Additionally, generalisability can be further divided into:

- Within-setting generalisability, where both training and testing data come from the same environment, such as in-lab images.
- Cross-setting generalisability, where the model is trained in one environment but evaluated in a different one, such as training on in-lab images and testing on in-the-wild images.

By considering both traditional evaluation metrics and generalisability factors, FER models can be assessed more holistically, ensuring they are not only accurate but also robust across different real world conditions.

D. FER Ethical Issues

In addition to going over research papers, dataset characteristics and model architectures its crucial to consider the ethical aspect of FER models. In accordance with the EU AI Act as outlined in [16] emotional recognition technology poses significant ethical challenges, particularly regarding privacy, consent, and potential misuse. Considering FER systems in real world environments they are often deployed without informed consent, especially in public or semi public contexts like workplaces, schools, or law enforcement, where per-individual consent is near impossible to achieve and power imbalances further complicate voluntary participation. Additionally the majority of FER systems have questionable accuracy when applied in real world scenarios primarily when dealing with non-white male individuals. This can further entrench discrimination, particularly when used to draw inferences about psychological states or even political beliefs, leading to profiling and unjust decision making. There are also broader societal harms, including distributive injustice where marginalised groups may be denied access to services and recognitional injustice, where identities and emotions are misinterpreted or dismissed. These harms are intensified by the opacity of AI systems, which can obscure how decisions are made and whether they are based on biased or flawed interpretations of emotional data.

In light of the above the EU Artificial Intelligence Act establishes a comprehensive, risk based legal framework to ensure trustworthy and human centric AI development. FER is explicitly addressed under the Act, which classifies its use in certain settings such as education and workplaces as unacceptable risk, and thus bans such applications outright. The Act also prohibits emotion recognition used for manipulation, exploitation of vulnerabilities, and untargeted biometric scraping for database expansion. In other contexts, FER may be classified as high risk, requiring strict compliance measures, including robust risk assessments, transparency, high quality datasets to mitigate bias, human oversight, and detailed documentation to ensure accountability. These provisions are part of a broader EU strategy to safeguard fundamental rights, prevent discrimination, and ensure that AI systems including FER do not undermine democratic principles or individual dignity [17].

Methodology and Evaluation

Following the literature review, the aforementioned five SOTA deep learning models were implemented for comparison. These models were sourced from Papers with Code [18], a platform widely used for sharing and benchmarking machine learning research. The corresponding source code for these models was retrieved primarily from GitHub repositories [19–22] and modified to ensure compatibility with the evaluation framework used in this study. Most models were originally tailored for specific datasets, necessitating adaptation to enable processing of other datasets.

Similarly the five aforementioned benchmark datasets were used in this study: AffectNet, RAF-DB, FER2013, JAFFE, and

CK+. Due to restricted access, complete versions of AffectNet and RAF-DB were unavailable, so Kaggle hosted subsets were utilised instead [23, 24]. Each dataset underwent preprocessing to conform to the input requirements of the models. This included reorganising dataset structures into flat directories with a corresponding label CSV file (`filename,label`) and into class specific subdirectories, depending on the model specifications. To ensure a consistent comparison across all datasets, the emotion labels were mapped to indices ranging from 0 to 7, as follows: 0: Angry, 1: Disgust, 2: Fear, 3: Happy, 4: Sad, 5: Surprise, 6: Neutral, and 7: Contempt.

Furthermore, datasets were split into training, validation, and test sets using predefined splits where available. Where such splits were not provided, new ones were created to ensure consistency across experiments. All images underwent facial feature alignment and cropping to standardise inputs. However, greyscale conversion was not applied uniformly: while some models relied on colour information, others included built in greyscale preprocessing, which was retained to allow each model to perform optimally. Notably, model performance across all datasets indicated that the absence of a global greyscale conversion had a negligible impact.

All model codes, excluding EmoNeXt, were modified to enhance clarity and remove redundant components present in their original repositories. These modifications were carefully validated to ensure that model performance remained consistent throughout. EmoNeXt was excluded from these changes due to its reliance on the `wandb` library for metric tracking, where any alteration to the codebase could compromise the accuracy of performance reporting. Retaining its original implementation ensured stability. Overall, these code refinements not only improved compatibility across datasets but also facilitated a deeper understanding of each model's architecture.

Furthermore, each model was trained using a standardised set of hyperparameters: a learning rate of 0.001, batch size of 16, momentum of 0.9, weight decay of 0.0001, a patience value of 15, and a maximum of 300 training epochs. Exceptions were made for EmoNeXt, which used a reduced learning rate of 0.0001 and an increased batch size of 32 to facilitate optimal performance. A consistent early stopping mechanism (patience = 15) was applied across all models to halt training when validation performance plateaued.

The evaluation strategy mirrored that outlined in [3]. Each model was trained on one of three in-the-wild datasets (FER2013, RAF-DB, AffectNet), which were selected for their size and diversity, offering a stronger foundation for generalisation compared to smaller, in-lab datasets (JAFFE and CK+). The latter were reserved for cross domain testing due to their limited sample sizes.

Given that the datasets do not all share the same set of emotion labels, special care was taken during evaluation. If a test image had a label that was not part of the model's training label set, it was excluded from evaluation. Conversely, if a model produced a prediction that was not present in the test set's label space, the prediction was considered invalid and the next most confident valid prediction was used instead.

Analysing the accuracy metrics presented in Tables I through V reveals that DDAMFN++ consistently outperformed all other models in overall accuracy, particularly when trained on RAF-DB, as shown in Table II. For instance, DDAMFN++ achieved an impressive 88.8% accuracy within RAF-DB and showed strong cross domain performance, such as 82.8% on CK+ when trained on RAF-DB. However, its performance dropped on the JAFFE dataset, where it consistently plateaued at 36.4% across all training sources. This suggests a lack of generalisation to JAFFE, likely due to its narrow demographic representation, in contrast to more diverse datasets like CK+. This discrepancy between JAFFE and CK+ performance is a recurrent trend across all models, indicating a dataset related bias that challenges cross domain generalisation, reinforcing the hypothesis that demographic diversity plays a crucial role in generalisation. Additionally raising valid concern regarding the ethical implications of the use of such models in real world scenarios.

Beyond classification accuracy, practical considerations such as model complexity and training efficiency are vital in real world deployment. As detailed in Table VI, ResNet based models (ResNet50 and ResEmoteNet) required significantly less training time across all datasets typically under 2.5 hours whilst still reaching competitive accuracy levels. For instance, ResNet50 trained on RAF-DB achieved 83.3% accuracy in just one hour, compared to DDAMFN++'s 88.8% in four hours. This highlights the efficiency advantage of ResNet variants for use cases with constrained computational resources and time.

Moreover, the RAF-DB dataset emerged as the most time efficient training source, consistently resulting in shorter training durations and high accuracy across models, likely attributed to its compact yet representative emotion class distribution.

Furthermore, model consistency and result replicability are critical for evaluating the reliability of any SOTA model. Table VII presents the accuracy metrics reported by Papers With Code, emphasising the differences between these results and those found in this research paper. Notably, the PAtt-Lite model shows a significant drop in performance between FER2013 and FER+, despite both datasets containing the same images with differing labels. A similar performance decline is observed for DDAMFN++ across these datasets, though the drop is less pronounced than in PAtt-Lite, suggesting that both models are sensitive to label distributions, with PAtt-Lite being more affected. Additionally, EmoNeXt and ResEmoteNet exhibit performance drops relative to their reported benchmarks, indicating a dependency on favourable training and testing splits. In contrast, excluding its FER2013 results, DDAMFN++ achieves performance that closely aligns with its reported accuracy, demonstrating stronger consistency and robustness. Nonetheless, because FER2013 and FER+ differ in their annotation schemes, they are not directly comparable, and any performance differences should be interpreted with caution.

Lastly, observing these results from an ethical lens it is clear as to why the EU AI act instilled regulations on the deployment of such models. In general the accuracies observed generally range from the low 50s to the high 70s across all scenarios. In cases of real world applications we are more concerned with the overall cross domain results as these best outline model generalisability, however these are less than satisfactory. Furthermore the significant dip in performance when testing on the JAFFE dataset further supports this claim as these models require a high degree of generalisability to be viable in real world applications.

In summary, DDAMFN++ trained on RAF-DB achieved the highest overall accuracy and showed strong generalisation within and across domains, making it the best model in terms of classification performance. This is mainly due to its attention mechanisms, which help the network focus on the most important facial features. Its use of feature mixing, coordinate attention, and an additional attention loss leads to better and more diverse feature learning. RAF-DB proved to be the best training dataset overall, offering strong performance across most models. On the other hand, JAFFE was the most useful for testing cross domain generalisation in niche or challenging scenarios. For situations where speed and efficiency matter, ResNet models like ResNet50 offer a good balance. They performed well while training much faster, thanks to their skip connections and efficient design that reduce complexity without sacrificing accuracy.

TABLE I
PAtt-Lite: Within & Cross-Setting Accuracy (%)

Train \ Test	AffectNet	RAF-DB	FER2013	JAFFE	CK+
AffectNet	**51.8**	60.0	40.5	41.0	61.1
RAF-DB	35.8	**80.2**	39.4	47.7	67.7
FER2013	31.8	55.7	**60.9**	25.0	74.2

TABLE II
DDAMFN++: Within & Cross-Setting Accuracy (%)

Train \ Test	AffectNet	RAF-DB	FER2013	JAFFE	CK+
AffectNet	**61.0**	68.5	49.7	36.4	75.8
RAF-DB	46.9	**88.8**	53.1	36.4	82.8
FER2013	38.0	45.8	**70.1**	36.4	81.7

TABLE III
EmoNeXt: Within & Cross-Setting Accuracy (%)

Train \ Test	AffectNet	RAF-DB	FER2013	JAFFE	CK+
AffectNet	**58.7**	63.5	44.3	36.4	70.0
RAF-DB	41.1	**78.0**	47.0	43.2	73.7
FER2013	43.0	64.1	**68.0**	22.7	80.6

TABLE IV
ResNet50: Within & Cross-Setting Accuracy (%)

Train \ Test	AffectNet	RAF-DB	FER2013	JAFFE	CK+
AffectNet	**56.0**	53.3	40.1	31.8	46.3
RAF-DB	43.7	**83.3**	49.5	34.1	79.6
FER2013	39.1	63.9	**67.3**	27.2	68.8

TABLE V
ResEmoteNet: Within & Cross-Setting Accuracy (%)

Train \ Test	AffectNet	RAF-DB	FER2013	JAFFE	CK+
AffectNet	**48.4**	60.3	43.7	38.6	74.7
RAF-DB	32.2	**78.1**	36.0	45.5	45.1
FER2013	35.8	58.0	**62.4**	47.7	66.7

TABLE VI
Model Training Time (in hours) and Corresponding Epochs (in brackets) Across Different Datasets

Dataset	Patt-Lite	DDAMFN++	EmoNeXt	ResNet50	ResEmoteNet
AffectNet	14h (111)	6.5h (34)	15h (99)	2h (19)	2h (47)
RAF-DB	2h (63)	4h (50)	3.5h (81)	1h (51)	1h (81)
FER2013	11h (132)	11h (68)	15h (145)	2.5h (50)	2h (76)

TABLE VII
Papers With Code Reported Model Accuracy (%)

Dataset	Patt-Lite	DDAMFN++	EmoNeXt	ResNet50	ResEmoteNet
AffectNet	-	65.04	64.13	-	72.93
RAF-DB	-	91.35	-	-	94.76
FER2013	-	-	-	-	79.79
FER+	95.55	90.74	-	-	-

Conclusion

This paper presented a comprehensive evaluation of five state-of-the-art FER models DDAMFN++, EmoNeXt, PAtt-Lite, ResNet50, and ResEmoteNet across five benchmark datasets: AffectNet, RAF-DB, FER2013, JAFFE, and CK+. The literature review established a foundational understanding of FER challenges, model architectures, and dataset characteristics, along with a discussion of the ethical considerations surrounding their real world deployment.

Through cross-domain and within-dataset experiments, DDAMFN++ emerged as the top performing model in terms of classification accuracy, particularly when trained on RAF-DB. On the other hand, ResNet based models such as ResNet50 offered a favourable trade off between performance and training efficiency, making them well suited for time or resource constrained applications.

However, the study also revealed significant limitations in the generalisability of all evaluated models especially on the JAFFE dataset highlighting a consistent performance drop when tested on data with limited demographic diversity. These findings reinforce ongoing concerns regarding bias and the ethical deployment of FER systems, particularly in high stakes or diverse real world environments.

In conclusion, while recent FER models show promising results under controlled or in distribution settings, their cross-domain robustness remains insufficient for reliable real world deployment. Future research should focus on improving generalisability through techniques such as domain adaptation, fairness aware training, and more inclusive dataset collection.

GitHub

The GitHub repository containing the code relevant to this project can be accessed via this link: GitHub link

References

[1] European Data Protection Supervisor, *TechDispatch: Facial Emotion Recognition*, Accessed: 2025-03-01, May 2021. DOI: 10.2804/519064. [Online]. Available: https://www.edps.europa.eu/system/files/2021-05/21-05-26_techdispatch-facial-emotion-recognition_ref_en.pdf.

[2] A. Rehman, M. Mujahid, A. Elyassih, B. AlGhofaily, and S. A. O. Bahaj, "Comprehensive review and analysis on facial emotion recognition: Performance insights into deep and traditional learning with current updates and challenges," *Computers, Materials & Continua*, vol. 82, no. 1, pp. 41–72, 2025, ISSN: 1546-2226. DOI: 10.32604/cmc.2024.058036. [Online]. Available: http://www.techscience.com/cmc/v82n1/59239.

[3] V. Suresh, G. Yeo, and D. C. Ong, *Critically examining the domain generalizability of facial expression recognition models*, 2023. arXiv: 2106.15453 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2106.15453.

[4] Goodfellow, I. J., D. Erhan, A. Courville, *et al.*, "Challenges in representation learning: A report on three machine learning contests," *arXiv preprint arXiv:1307.0414*, 2013.

[5] S. Li, W. Deng, and J. Du, *RAF-DB: Real-world Affective Faces Database*, Website, Accessed: 2025-03-01, 2017. [Online]. Available: http://www.whdeng.cn/RAF/model1.html#dataset.

[6] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computation in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017. DOI: 10.1109/TAFFC.2017.2740923.

[7] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. 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*, 2010, pp. 94–101. DOI: 10.1109/CVPRW.2010.5543262.

[8] M. J. Lyons, *"excavating ai" re-excavated: Debunking a fallacious account of the jaffe dataset*, Jul. 2021. DOI: 10.5281/zenodo.5147170. [Online]. Available: https://doi.org/10.5281/zenodo.5147170.

[9] M. Lyons, M. Kamachi, and J. Gyoba, *The japanese female facial expression (jaffe) dataset*, Zenodo, Apr. 1998. DOI: 10.5281/zenodo.3451524. [Online]. Available: https://doi.org/10.5281/zenodo.3451524.

[10] Y. El Boudouri and A. Bohi, "Emonext: An adapted convnext for facial emotion recognition," in *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, 2023, pp. 1–6.

[11] A. K. Roy, H. K. Kathania, A. Sharma, A. Dey, and M. S. A. Ansari, *Resemotenet: Bridging accuracy and loss reduction in facial emotion recognition*, 2024. arXiv: 2409.10545 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2409.10545.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] *Explicit content detection system: An approach towards a safe and ethical environment - scientific figure on researchgate*, https://www.researchgate.net/figure/Block-diagram-of-Resnet-50-1-by-2-architecture_fig4_326198791, [Accessed: 22 Mar 2025], 2025.

[14] S. Zhang, Y. Zhang, Y. Zhang, Y. Wang, and Z. Song, "A dual-direction attention mixed feature network for facial expression recognition," *Electronics*, vol. 12, no. 17, 2023, ISSN: 2079-9292. DOI: 10.3390/electronics12173595. [Online]. Available: https://www.mdpi.com/2079-9292/12/17/3595.

[15] J. Le Ngwe, K. M. Lim, C. P. Lee, T. S. Ong, and A. Alqahtani, "Patt-lite: Lightweight patch and attention mobilenet for challenging facial expression recognition," *IEEE Access*, 2024.

[16] E. Parliament, D.-G. for Parliamentary Research Services, T. Madiega, and H. Mildebrath, *Regulating facial recognition in the EU – In-depth analysis*. European Parliament, 2021. DOI: doi/10.2861/140928.

[17] E. Commission, *Ai act — shaping europe's digital future*, Accessed: 2025-05-02, 2025. [Online]. Available: https://digital-trategy.ec.europa.eu/en/policies/regulatory-framework-ai.

[18] *Papers with code (fer)*, https://paperswithcode.com/task/facial-expression-recognition, Accessed: March 14, 2025.

[19] S. Zhang, *Ddamfn: A dual-direction attention mixed feature network for facial expression recognition*, https://github.com/SainingZhang/DDAMFN, 2025.

[20] Y. E. Boudouri and A. Bohi, *Emonext: An adapted convnext for facial emotion recognition*, 2025. [Online]. Available: https://github.com/yelboudouri/EmoNeXt.

[21] J. L. Ngwe, K. M. Lim, C. P. Lee, T. S. Ong, and A. Alqahtani, *Patt-lite: Lightweight patch and attention mobilenet for challenging facial expression recognition*, 2024. [Online]. Available: https://github.com/JLREx/PAtt-Lite.

[22] A. K. Roy, H. K. Kathania, A. Sharma, A. Dey, and M. S. A. Ansari, *Resemotenet: Bridging accuracy and loss reduction in facial emotion recognition*, Accessed: 2025-04-24, 2024. [Online]. Available: https://github.com/ArnabKumarRoy02/ResEmoteNet.

[23] thienkhonghoc, *AffectNet Dataset (Unofficial)*, Kaggle Dataset, Accessed: 2023-10-05, 2023. [Online]. Available: https://www.kaggle.com/datasets/thienkhonghoc/affectnet.

[24] shuvoalok, *Unofficial Mirror of RAF-DB Dataset*, Kaggle Dataset, Accessed: 2025-03-01, 2023. [Online]. Available: https://www.kaggle.com/datasets/shuvoalok/raf-db-dataset.