IEEE Open Journal of Signal Processing

Received 13 September 2024; revised 10 December 2024; accepted 19 December 2024. Date of publication 16 January 2025; date of current version 27 February 2025. The review of this article was arranged by Associate Editor L. A. Thomaz.

Digital Object Identifier 10.1109/OJSP.2025.3530793

Mixture of Emotion Dependent Experts: Facial Expressions Recognition in Videos Through Stacked Expert Models

ALI N. SALMAN [1] (Member, IEEE), KAREN ROSERO [2] (Member, IEEE), LUCAS GONCALVES [1] (Graduate Student Member, IEEE), AND CARLOS BUSSO [1,2] (Fellow, IEEE)

[1]The University of Texas at Dallas, Richardson, TX 75080 USA
[2]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA

CORRESPONDING AUTHOR: CARLOS BUSSO (e-mail: busso@cmu.edu).

ABSTRACT Recent advancements in *dynamic facial expression recognition* (DFER) have predominantly utilized static features, which are theoretically inferior to dynamic features. However, models fully trained with dynamic features often suffer from over-fitting due to the limited size and diversity of the training data for fully *supervised learning* (SL) models. A significant challenge with existing models based on static features in recognizing emotions from videos is their tendency to form biased representations, often unbalanced or skewed towards more prevalent or basic emotional features present in the static domain, especially with posed expression. Therefore, this approach under-represents the nuances present in the dynamic domain. To address this issue, our study introduces a novel approach that we refer to as *mixture of emotion-dependent experts* (MoEDE). This strategy relies on emotion-specific feature extractors to produce more diverse emotional static features to train DFER systems. Each emotion-dependent expert focuses exclusively on one emotional category, formulating the problem as binary classifiers. Our DFER model combines these static representations with recurrent models, modeling their temporal relationships. The proposed MoEDE DFER approach achieves a macro F1-score of 74.5% , marking a significant improvement over the baseline, which presented a macro F1-score of 70.9% . The DFER baseline is similar to MoEDE, but it uses a single static feature extractor rather than stacked extractors. Additionally, our proposed approach shows consistent improvements compared to other four popular baselines.

INDEX TERMS Video facial expression recognition, emotion recognition, affective computing, ensemble model.

I. INTRODUCTION

Emotions play an important role in interpersonal communication, where human facial expressions convey rich nonverbal cues filled with emotional content. The automatic detection of emotions from facial features opens a broad range of applications in affective computing. These applications include enhancing *human-computer interaction* (HCI), improving security and surveillance systems, monitoring mental health conditions, and evaluating engagement levels in educational settings [7]. The field of *facial expression recognition* (FER) is a challenging task due to the intricate and diverse expressions exhibited across human faces. Studies have explored FER solutions using both static and dynamic strategies. Systems that operate with static images are termed *static facial expression recognition* (SFER) systems, while those analyzing video sequences to infer the underlying emotions are referred to as *dynamic facial expression recognition* (DFER) systems. While considerable progress has been made in discriminating posed expressions in static images, DFER remains a major challenge, particularly in the presence of speech where facial movements due to speech articulations add variability [23], [33]. Our project primarily focuses on enhancing DFER performance.

The significance of feature representations in FER cannot be overstated, as they serve as the basis for effectively capturing and distinguishing different emotional expressions

in dynamic facial images or video sequences. Deep learning models designed for SFER have demonstrated competitive features, particularly when trained on extensive datasets encompassing facial images from diverse subjects, ages, ethnicities, illumination, background, angles, and environments, such as the AffectNet corpus [31]. However, despite the huge size of the AffectNet corpus (approximately 440,000 labeled images), the distribution of images across its eight discrete emotional labels is unbalanced, with classes like disgust and contempt being notably underrepresented. Consequently, static deep feature extractors may inadequately represent and exhibit less efficacy, especially for underrepresented emotions. Moreover, while feature representations derived from SFER systems are valuable, they may require fine-tuning before being employed for DFER [14], [27], [32], [37]. This need arises from the direct impact of speech articulation on facial expressions [3], [4], particularly in the movements of the orofacial region, which may not align seamlessly with expressions characterized by static feature extractors. Additionally, feature extractors trained exclusively in the dynamic domain are susceptible to rapid over-fitting due to the limited size and diversity of the training data, especially when compared to static datasets. As a result, they typically require pre-training on large unlabeled video-based datasets using self-supervised learning methods, which is time-consuming and necessitates using a GPU cluster [38]. Therefore, it is important to develop a pre-training approach that not only produces diverse and balanced static feature representations that can be fine-tuned for the dynamic domain, improving FER performance in video sequences but can also be trained on widely available consumer-grade GPUs within hours.

Given the identified limitations, this study aims to enhance the diversity and generalizability of static feature extractors, improving the performance of dynamic FER. We hypothesize that a singular feature vector representing balanced or unbalanced classes may lack the necessary diversity, as it is an average representation of all emotions considered in the dataset. To address this limitation, we propose a *mixture of emotion dependent experts* (MoEDE) strategy, which trains individual 'expert' feature extractors for each emotional category. The objective is for each emotion-based expert feature extractor to learn discriminative features specific to its corresponding emotion, facilitating the distinction of that emotion from others within the dataset. Since our main focus is on DFER, we extend the application of each expert feature extractor to the dynamic domain. We transfer the static features to the dynamic domain by incorporating a *long short-term memory* (LSTM) layer. This layer learns the temporal dependencies within the emotion-dependent expert input sequence, followed by a classification layer that independently predicts the emotion in an image sequence using each emotional expert. Concurrently, during the same training process, we employ an additional LSTM layer to ensemble all dynamic outputs resulting from the previous LSTMs of the expert feature extractors, followed by a final softmax layer to predict the underlying emotion in the image sequence using all emotion-dependent experts.

We train eight emotion-dependent expert feature extractors for SFER using binary classification. For this purpose, we used subsets of the AffectNet dataset specific to each expert. Each subset includes all samples corresponding to the expert's positive emotion and an equivalent number of randomly selected samples from the remaining emotional classes (e.g., happiness versus selected samples not from the class happiness). We train and evaluate our MoEDE framework for DFER using the CREMA-D corpus [5]. Our experiments demonstrate that by training on positive emotions, we can achieve an average F1-score of 84.13% on just the positive emotions in the development set of the AffectNet corpus. Additionally, when analyzing the features extracted by each emotion-dependent expert using the cosine similarity, we found that the average cosine similarity across all feature extractors is 0.33. This result indicates that while there are similarities among the features, they can significantly differ. We train our DFER MoEDE model by employing the emotion-specific models, obtaining macro F1-scores that outperform all other baseline models. The contributions of this paper are as follows:

- A set of emotion-dependent expert feature extractors aimed at enhancing the diversity and generalization of feature representations in SFER.
- An ensemble model comprising individual expert classifiers demonstrating superior performance compared to existing systems for DFER.

We are releasing the code and MoEDE models for public use.[1]

II. RELATED WORK

The following section provides an overview of current static and dynamic FER advancements. Subsequently, it examines SFER techniques, which aim to enrich static feature diversity. Furthermore, it discusses relevant ensemble models employed for DFER.

A. FACIAL EXPRESSION RECOGNITION

We present an overview of deep learning methodologies designed for FER based on both static images and dynamic sequences. FER on static images has achieved impressive performance across diverse datasets and under real-world conditions [30]. For instance, Zhang et al. [42] proposed a mixed feature network with a dual-direction attention head to generate attention maps based on the extracted features from various facial regions. Mao et al. [28] combined multi-scale features computed via a window-based cross-attention mechanism and facial landmarks for improved static FER. Additionally, Savchenko et al. [35] proposed a lightweight yet competitive FER model based on the EfficientNet [39] and MobileNet [19] architectures, suitable for devices lacking powerful GPUs. Li et al. [26] employed an encoder-decoder

[1] [Online]. Available: https://github.com/3loi/MoEDE

architecture to map facial expressions to a common representation, decompose it into emotional and orthogonal components, and synthesize emotionless faces, ultimately enabling emotion recognition.

Research indicates that even short image sequences depicting emotional dynamism facilitate a more accurate and consistent perception of emotions [8]. However, despite the advantages of analyzing facial expressions within dynamic video sequences, factors such as speech content, cultural nuances, physiological differences, and speaker idiosyncrasies can affect the performance of FER systems [29]. To address these issues, sophisticated techniques are required to delineate the emotion-dependent correlation between facial gestures and speech [4], given that speech articulation continuously alters facial appearance [3]. Consequently, *state-of-the-art* (SOTA) approaches in dynamic emotion recognition have explored multimodal frameworks, encompassing both audio and visual modalities. Since our study exclusively addresses the visual modality, we review recent audiovisual approaches that have reported their efficacy when solely relying on visual modality.

Huang et al. [21] proposed a model that employs a transformer-based architecture [41] for feature extraction, followed by an LSTM layer, aimed at improving the modeling of temporal dependencies. We refer to this method as the T-LSTM framework. Goncalves and Busso [12], [13] proposed the AuxFormer, an audio-visual transformer-based framework featuring unimodal auxiliary networks, enabling the evaluation of performance on individual modalities. Gong et al. [15] proposed the *unified audio-visual model* (UAVM) that integrates the transformer and classification layers shared across modalities, accompanied by modality-specific feature extractors. Lastly, Goncalves et al. [14] introduced the *versatile audio-visual learning* (VAVL) method, which is a versatile audio-visual learning framework capable of accommodating single or multiple modalities by using shared layers used by all the modalities. For the visual modality, they used a pre-trained EfficientNet-B2 model [35] to extract feature representations from input faces, followed by a conformer encoder [16] to process the video sequence features.

This study explores the potential of diverse and enriched feature representations extracted from the static domain to enhance FER in video sequences. Unlike previous studies, we adopt an ensemble approach, using multiple emotion-dependent expert models trained for each emotional class for improved DFER.

B. FEATURE DIVERSIFICATION FOR STATIC FER

In previous studies within the field of emotion recognition, a persistent issue has been the lack of diversified feature representations. A common approach for addressing this limitation involves training deep learning models with different architectures. This strategy aims to enhance feature generalization, consequently improving the performance of a general emotion recognition classifier beyond that of individual networks focused on the same task. Jia et al. [22] trained well-established architectures based on *convolutional neural networks* (CNNs) to extract features from images portraying emotions including AlexNet [25] and ResNet [18]. Subsequently, the diversified features were combined to train a *support vector machine* (SVM) for the final emotion categorization. The authors demonstrated the superiority of the SVM classifier over training the individual models for the same task. Similarly, Agustin et al. [1] independently trained CNN models including MobileNet [19] for SFER. Features derived from these pre-trained models were concatenated. A dense layer was then used to reduce its dimension. The classifier trained with the concatenated diversified features outperformed each individual model. Siqueira et al. [36] explored the use of four identical branches of the same CNN model, all trained for FER. Diversified features were extracted from different layers of each model. The heatmaps were employed to illustrate that the features analyze different facial regions to determine an emotion. While the generalization of these static features has been improved for SFER, they do not necessarily imply improvement for DFER. In contrast, a key concern addressed in our work is expanding the representation diversity of static features to enhance FER performance in videos (e.g., dynamic domain).

C. ENSEMBLE MODELS FOR DYNAMIC FER

Accurate dynamic FER requires features that adeptly capture subtle facial movements associated with emotions, which can also be influenced by speech. Traditional methodologies have relied on handcrafted features. For instance, Chen et al. [6] utilized a histogram of oriented gradients to extract dynamic textures from video sequences alongside a geometric feature derived from facial landmark warp transformations. These features were then processed using a multiple kernel SVM approach to predict the class for the entire sequence. However, while features were designed to encapsulate the temporal dynamics of facial expressions in videos, the multiple kernel SVM may struggle to capture intricate temporal patterns within the video sequence. Additionally, researchers have explored the combination of manually engineered features and those obtained from deep learning models to characterize facial appearance changes in video sequences. Hao et al. [17] combined two classifiers; the first is an SVM that uses texture deformation features known as *local binary patterns* (LBP) to extract hand-crafted features, and the second is CNN-based model inspired by AlexNet that extract deep features directly from the images. Both of these classifiers are combined using decision-level fusion during prediction. Similarly, Do et al. [10] introduced an ensemble of three distinct CNN-based face representation models, each incorporating a module to capture sequential information from videos. Despite the effectiveness of employing a weighted fusion approach, emotion recognition accuracy for short videos reached values of around 55% . While prior research has predominantly focused on diverse facial regions and varied deep-learning architectures, an unexplored avenue remains: extracting emotion-specific features. We aim to harness the

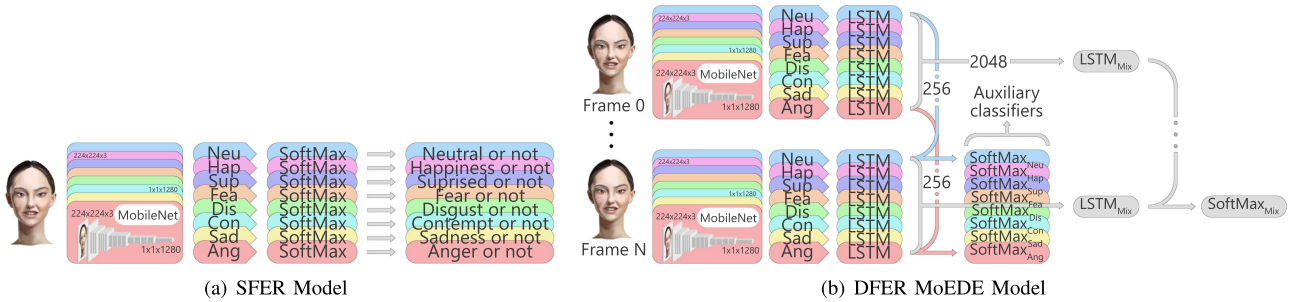
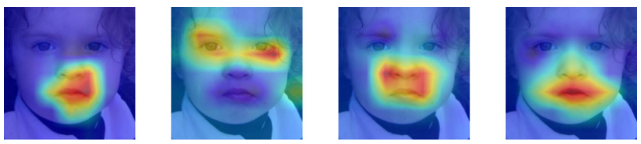


FIGURE 1. Diagram of the proposed FER system. The model consists of two parts (a) the static image predictor model, which predicts the facial expression for a single face/image, and (b) a dynamic video predictor, which aggregates the high-level features extracted from the image predictor to recognize emotions in the video.



(a) Anger (b) Sadness (c) Fear (d) Disgust

FIGURE 2. Shows the CAM [43] localization applied on the global pooling layer of MobileNetV2 on different emotion-specific models, on a sample image in the AffectNet development set.

potential of ensemble techniques and compare them against current SOTA methodologies for DFER.

III. METHODOLOGY

Our proposed approach aims to diversify the facial features extracted in the static domain to improve the performance of categorizing emotions in the dynamic domain. As such, we split our methodology into two parts: (1) training the static emotion classifiers (Section A), and (2) training the dynamic emotion classifiers (Section B).

A. STATIC FACIAL EXPRESSION RECOGNITION

A common approach for SFER is to train a single model to differentiate between all categorical emotions. This general SFER model is represented as $SFER(x_i) = g(f(x_i))$, where an image x_i is processed through a feature extractor $f(x_i)$ and a static classifier $g(\cdot)$ to predict the emotion. This model provides a general space of features extracted from emotional facial images. However, it may develop biases toward more prevalent or easily recognizable emotions.

Our proposed approach aims to enhance DFER performance by diversifying the static-level feature representation $f(x_i)$, for which we train individual emotion-dependent feature extractors, as seen in Fig. 1(a), denoted as $SFER_{emo}(x_i) = g_{emo}(f_{emo}(x_i))$. We use the pre-trained general SFER model as a starting point to finetune each emotion-dependent SFER as a binary classification. This process adjusts the feature space to be more focused on specific emotions. We train 8 separate models, each specializing in one emotion.

For instance, $SFER_{hap}(x_i)$ classifies the image as either happy or not happy. We select all the samples from the target emotions in the train set for the positive class. We select all samples from the target class and an equal-sized random set from other emotions to ensure balanced sets for each training epoch. This practice mitigates sample bias toward the most represented classes, and along with traditional image data augmentation approaches, reduces the risk of overfitting.

By isolating the feature space for each emotion, these extractors learn fine-grained, emotion-specific characteristics, allowing the model to capture subtle variations that might be overlooked by a single general extractor. This specialized approach enhances the model's ability to recognize emotions effectively from the diverse feature space extracted from all the emotion-specific feature extractors.

In total, we train eight different emotional classifiers. The outputs from all feature extractors, $f_{emo}(\cdot)$, are aggregated to represent the MoEDE features $f_{MoEDE}(\cdot)$, where $f_{emo}(\cdot) \in \mathbb{R}^{1,280}$ and $f_{MoEDE}(\cdot) \in \mathbb{R}^{8\times1,280}$. However, as we employ eight different models to extract features for a single frame, we must use small, efficient models. We have chosen the MobileNetV2 model [34] for this purpose, which is a lightweight CNN-based model with just over 2 million parameters per model, totaling just over 17 million parameters across the eight SFER models. The total number of parameters is still less than those in other networks, such as the ResNet-50 model (around 25 million parameters [18]). that we use the standard MobileNetV2 architecture, changing only the 1000-class classifier output into a 2-class classifier.

B. DYNAMIC FACIAL EXPRESSION RECOGNITION

We freeze the SFER models to extract emotion-dependent features for training a DFER model. Freezing the weights is crucial as it prevents the model from overfitting to the video dataset's nuances, leveraging instead the generalization capabilities developed during initial training.

As shown in Fig. 1(b), our proposed architecture incorporates two layers of LSTM networks. The first layer contains 8 LSTMs, which separately process the features from each emotion-dependent extractor. The 1,280D output vector of each feature extractor $f_{emo}(\cdot)$ serves as

input to an individual LSTM layer, producing a 256-dimension output vector for each emotion feature extractor $LSTM_{emo}(f_{emo}(x_1), f_{emo}(x_2), f_{emo}(x_3), \ldots)$, where $x = x_1, x_2, \ldots$ represents the image sequence from a video.

These 256-dimension vectors are then concatenated along the time axis, forming a 2,048-dimension vector (8×256) that represents the aggregated features for each frame. This vector feeds into another LSTM, referred to as $LSTM_{mix}$, which reduces it to a 256-dimension feature vector. It is important to note that the different feature extractors' outputs are combined only in the $LSTM_{mix}$; all previous operations are independently conducted, enabling parallel computing if needed.

In our initial tests, using a softmax classifier solely at the last timestep of the $LSTM_{mix}$ led to the model predominantly backpropagating into only a few of the emotion-specific LSTMs ($LSTM_{emo}$). To ensure that the backpropagation impacts all the components of the model, we incorporate a softmax classifier at the last timestep of each emotion-specific LSTM as auxiliary tasks. This modification aims to improve learning across all network parts by independently providing gradient updates to each emotion-specific LSTM layer. Note that the auxiliary classifiers in the DFER implementation predict all the emotional classes, not just binary.

IV. RESOURCES AND IMPLEMENTATION

A. RESOURCES

1) AFFECTNET [31]

This comprehensive facial expression database contains over 1 million images sourced from the internet. Unlike video-based datasets, which are limited in the number of subjects, this corpus includes approximately 440,000 manually annotated images, which include thousands of different individuals. Each image is categorized into one of eight discrete emotional categories. Although the annotations include emotional attributes such as valence and arousal, our study focuses exclusively on these eight emotional labels. On average, the images in the AffectNet corpus are 425×425 pixels in resolution. To train our expert feature extractors, we employ a balanced sampling strategy. In each epoch, we select all samples from the positive emotion class and an equal number of images randomly sampled from the other emotions. We utilize the provided training and development sets included with the dataset and report our results on the development set, as the official test set is not publicly available.

2) CREMA-D [5]

The dataset is an audiovisual collection of videos featuring individuals demonstrating predefined emotional attributes such as happiness, fear, disgust, anger, sadness, and neutral state. It comprises recordings from 91 actors (48 male and 43 female) representing diverse ethnicities and age groups. Each actor was directed to express specific emotional states while pronouncing a target sentence. At least four raters have annotated every sample in the dataset across three different modalities: audio-only, video-only, and audio-visual. The dataset includes a total of 7,442 clips. We formulate the emotion recognition task as a six-class problem for the emotions happiness, fear, disgust, anger, sadness, and neutral state. The distribution of sentences per emotional class is not uniform. Recordings were made in a controlled environment with a resolution of 960×720 pixels against a green screen to ensure consistent frontal video capture. For our analysis, the consensus label is the majority rule from the audio-visual annotations. To provide a more accurate representation of our model's performance and further ensure robustness, we have structured the dataset into five different participant-independent splits, where each subject is in exactly one set. Each split consists of training (70%), development (15%), and testing (15%).

B. IMPLEMENTATION DETAILS

In all our experiments, we utilize the ADAM optimizer [11] with a learning rate of 0.001. For the SFER, we use the weighted categorical cross-entropy loss using the AffectNet corpus. For the DFER model, we employ the categorical cross-entropy loss on the CREMA-D corpus. We preprocess all facial images from the AffectNet and CREMA-D by extracting the facial region and landmarks using the dlib toolkit [24]. Using the landmarks, we further rotate the images so that the location of the eyes is consistent across all images and parallel with the x-axis, achieving a fronto-parallel view. The detected landmarks are used to crop the facial region, ensuring consistent and standardized facial images.

For the SFER model, we use the weights of the pretrained MobileNetV2 on ImageNet [9]. Starting from this model, we train an 8-class classifier on the AffectNet by swapping the classification head from 1000 classes to 8 classes. We train this model for 60 epochs and use it as the starting point for each emotion-dependent expert model. To train an emotion-dependent model, we change the classification head from 8 to 2 classes and fine-tune it using the aforementioned balanced emotion-dependent subset of the AffectNet for 20 epochs. Additionally, during training on AffectNet, we apply data augmentation techniques, including up to 15 degrees of random rotation, 10% of random translation, and 4% of random shear.

For the DFER MoEDE model, we freeze the CNN portion of the MobileNetV2 network and extract a feature vector with dimensions equal to 1,280 from each image's global pooling layer. These vectors, denoted as $x_v \in \mathbb{R}^{N_v \times 1,280}$, capture the emotional visual features, where N_v represents the total number of frames analyzed per sequence and 1,280 reflects the dimensional feature size extracted from each frame by the MobileNetV2 architecture. Additionally, we apply zero post-padding on each mini-batch based on the longest image sequence in the batch. Then, we train the DFER model, backpropagating the loss through the eight emotion-specific LSTMs ($LSTM_{emo}$), as well as the fusion LSTM ($LSTM_{mix}$). Using this approach, we train five models on the different splits of CREMA-D corpus. We train for 20 epochs for each model, saving the best model on the development set based on accuracy. All the code implemented is in Python and Pytorch

TABLE 1. Detailed Performance Metrics by Emotion (Percentage) for Each MobileNetv2 Model Trained on a Binary Problem: Positive or Target Emotion, and None Positive (negative) Emotion Samples. (P: Positive Emotion, N: Negative Emotions)

Emotion	P Precision [% ↑]	N Precision [% ↑]	P Recall [% ↑]	N Recall [% ↑]	P F1-score [% ↑]	N F1-score [% ↑]	F1-macro [% ↑]	F1-weighted [% ↑]	Accuracy [% ↑]
Neutral	87.98	13.46	66.74	36.20	75.91	19.62	47.76	68.87	62.93
Happiness	84.05	1.08	73.77	2.00	78.58	1.40	39.99	68.93	64.80
Sadness	87.21	11.32	79.86	18.00	83.37	13.90	48.64	74.69	72.13
Surprise	85.68	4.65	79.49	7.00	82.47	5.59	44.03	72.86	70.43
Fear	86.05	4.65	83.03	5.80	84.51	5.16	44.84	74.59	73.38
Disgust	87.71	14.43	90.51	11.20	89.09	12.61	50.85	79.53	80.60
Anger	95.65	42.66	86.06	72.60	90.60	53.74	72.17	85.99	84.38
Contempt	86.71	4.29	90.43	3.00	88.53	3.53	46.03	77.91	79.50
Average	87.63	12.07	81.24	19.47	84.13	14.44	49.29	75.42	73.52

and trained on a single RTX 4060, taking slightly over one day to train all five models, including the AffectNet pre-training (starting from the ImageNet weights), highlighting the training efficiency of our model. During evaluation, we report the mean metrics of the test sets across the five splits. Additionally, we assess statistical significance using a one-tailed matched-pair t-test, dividing each split into five sets for a total of 25 sets. Significance is determined at a p-value < 0.05.

C. COMPUTATIONAL COMPLEXITY

The complexity of our model is reported in terms of the number of parameters and computational requirements. Each feature extractor in the model is based on the MobileNetV2 architecture, comprising 2.2 million parameters. A total of 17.79 million parameters are optimized for the eight experts in our model. Additionally, the fusion model, which integrates the representations of these experts, contributes with 14.98 million parameters. Overall, the MoEDE framework comprises 32.76 million parameters, requiring approximately 122 MB of storage when using 32-bit floating-point precision.

In terms of computational demand, processing 2 seconds of video data (60 frames) with MoEDE requires approximately 0.32 teraFLOP (TFLOP). For reference, current smartphones such as the iPhone 15 Pro, can handle around 2.15 TFLOP per second.

Furthermore, we anticipate that MoEDE can benefit significantly from efficient parameter fine-tuning techniques, such as *low-rank adaptation* (LoRA) [20]. Since all eight experts are derived from a shared base model and fine-tuned for specific emotional tasks, employing low-rank weight updates could reduce the model size by over 7x.

V. EXPERIMENTAL RESULTS

This section evaluates our proposed SFER and DFER models. Initially, we report the F1-score and accuracy for our eight-class and two-class SFER models trained on the AffectNet corpus. We demonstrate the effectiveness and diversity of the different models by comparing the features extracted from each emotion-dependent expert. Next, we evaluate our dynamic models by analyzing the output from each emotion-specific feature extractor on the CREMA-D corpus. Finally, we evaluate the performance of our proposed MoEDE DFER model, comparing the results with four SOTA baselines.

TABLE 2. Feature Extractor Performance Implemented With the MobileNetV2, Trained on an 8 Class Problem on AffectNet. Metrics Show the AffectNet 8 Class-Balanced Validation Set

Emotion	Precision ↑ [%]	Recall ↑ [%]	F1-score ↑ [%]	Accuracy ↑ [%]
Neutral	48.8	58.0	53.0	58.0
Happiness	72.4	75.4	73.8	75.4
Sadness	65.3	62.0	63.6	62.0
Surprised	60.0	55.2	57.5	55.2
Fear	67.4	63.2	65.2	63.2
Disgust	58.7	62.6	60.6	62.6
Anger	59.8	55.4	57.5	55.4
Contempt	60.4	58.6	59.5	58.6
Average	61.6	61.3	61.3	61.3

A. SFER RESULTS

For the SFER, we first evaluate the performance of the AffectNet MobileNetV2 model trained on eight classes. As shown in Table 2, the model achieves a respectable 61.3% F1-score and 61.3% accuracy on the development set of the AffectNet corpus. This performance is comparable to the slightly larger EfficentNet-B0 model [40] trained by Savchenko et al. [35], which achieved an F1-score of 61.3% . Additionally, our model achieves a comparable score for both precision (61.6%) and recall (61.3%), showing that, overall, the model makes the same number of type 1 and type 2 errors. Furthermore, the table shows that even with the weighted categorical cross-entropy loss, some emotions achieve a high F1-score (e.g., happiness 73.8%), and some emotions achieve a low F1-score (anger 57.5%). Surprisingly, the class with the lowest F1-score is neutral, with an F1-score of 53.0% , while having the second highest number of training samples.

Next, we evaluate the binary emotion-dependent expert models. These models are binary classifiers, targeting a specific emotion as a positive sample (P) while treating all remaining samples as negative examples (N). As we explain in Section IV, we use the development set of the AffectNet corpus to evaluate our models. This set is balanced across emotions, including 500 samples per class. Therefore, the positive class has 500 samples, and the negative class has 3,500 samples.

Table 1 displays the evaluation metrics. We observe significantly higher F1-scores for the positive class (P) than the negative class (N). Specifically, the average F1-score for the

TABLE 3. MoEDE Feature Extractor Performance Implemented With the MobileNetV2 on the AffectNet Validation Set. We Perform Two Experiments. FFE: Frozen Feature Extractors. FTFE: Fine-Tuned Feature Extractors. (P: Precision Rate, R: Recall Rate, F1: Macro F1-Score, Ac: Accuracy)

Model	P↑ [%]	R↑ [%]	F1↑ [%]	Ac↑ [%]	P↑ [%]	R↑ [%]	F1↑ [%]	Ac↑ [%]
CNN	FFE				FTFE			
Neutral	41.1	61.2	49.2	61.2	46.8	58.0	51.8	58.0
Happiness	58.1	87.0	69.7	87.0	72.6	76.4	74.5	76.4
Sadness	59.7	62.0	60.8	62.0	63.1	61.8	62.4	61.8
Surprised	55.1	62.6	58.6	62.6	55.3	66.0	60.2	66.0
Fear	70.2	54.6	61.4	54.6	68.4	58.4	63.0	58.4
Disgust	76.2	37.8	50.5	37.8	70.9	47.2	56.7	47.2
Anger	53.3	58.6	55.8	58.6	55.5	60.4	57.9	60.4
Contempt	67.8	31.6	43.1	31.6	61.6	57.2	59.3	57.2
Average	60.2	56.9	56.1	56.9	61.8	60.7	60.7	60.7

TABLE 4. Cosine Similarity Between the Individual Expert Classifiers $SFER_{emo}(\cdot)$ on AffectNet Development Set

	Neu	Hap	Sad	Sur	Fea	Dis	Ang	Con
Neu	1.00	0.21	0.28	0.28	0.33	0.33	0.29	0.31
Hap	0.21	1.00	0.22	0.23	0.25	0.28	0.22	0.27
Sad	0.28	0.22	1.00	0.32	0.35	0.37	0.32	0.35
Sur	0.28	0.23	0.32	1.00	0.42	0.39	0.34	0.38
Fea	0.33	0.25	0.35	0.42	1.00	0.44	0.36	0.41
Dis	0.33	0.28	0.37	0.39	0.44	1.00	0.40	0.45
Ang	0.29	0.22	0.32	0.34	0.36	0.40	1.00	0.37
Con	0.31	0.27	0.35	0.38	0.41	0.45	0.37	1.00

positive class (P) across all emotions is 84.13% , and for the negative class (N) is only 14.44% . This disparity indicates a bias due to data imbalance, further emphasized by the differences in F1-macro and F1-weighted scores. The F1-macro score is 49.29% . This metric is a weighted average across emotions and is considerably lower than the unweighted F1-weighted score (75.42%). Individually, the F1-score for each targeted emotion is relatively high compared to the eight-class model results (see Table 2). For example, anger achieves an F1-score just above 90% , a significant improvement over the 57.5% observed in the eight-class model. Conversely, neutral records the lowest F1-score in both the binary and eight-class models, at 75.91% and 53.0% , respectively.

We want to compare two alternative implementations for fusing the eight emotion-dependent experts: *frozen feature extractors* (FFE), where the error does not propagate in the emotion-dependent feature extractors, and *fine-tuned feature extractors* (FTFE), where the error propagate in the emotion-dependent feature extractors. We evaluate these two implementations to predict the eight classes in the AffectNet corpus. Both models share the same architecture, which includes a fully connected layer that downsamples each 1,280-dimension feature into a 256-dimensional vector, followed by a concatenation of all features to form a 2048-dimension vector (8×256). The last layer is an eight-class softmax layer to predict emotions on the AffectNet corpus. Both models are trained similarly to the eight-class model described in Section III. Table 3 reports the evaluations of these models. We observe that the overall F1-score and accuracy of the fully trained model (FTFE) are significantly higher than those of the partly trained model (FFE). FTFE achieves overall improvements of 4.5% (absolute) in F1-score and 3.8% (absolute) in accuracy, respectively. This performance gain is attributed to the FTFE's ability to backpropagate the weighted loss through all the layers, unlike in the FFE model. However, for certain classes, such as happiness, neutral, and sadness, the FFE model shows higher accuracy, underscoring the presence of data imbalance, as these are the three classes with the most training samples.

B. DFER RESULTS

The first block of results from Table 5 illustrates the performance of different configurations of our MoEDE model. We highlight the impact of integrating auxiliary classifiers in emotion-specific LSTM layers (Section B), comparing Frozen Feature Extractors (FFE) and Fine-Tuned Feature Extractors (FTFE), as listed in Table 3. Our findings reveal that auxiliary classifiers boost the F1-weighted score to 78.9% for FFE (2.1% absolute increase) and to 73.5% for FTFE (1.6% absolute increase). Interestingly, despite lower performance on the AffectNet corpus, the model FFE achieves a higher F1-weighted score of 78.9% in dynamic settings, outperforming the results from the FTFE model (77.9%). This result underscores an important aspect: superior image-based classifiers do not necessarily translate to better video-based classifiers, reflecting the distinct and additional complexities inherent in video data.

The second block of results in Table 5 corresponds to the results obtained using the auxiliary classifiers. Our proposed MoEDE DFER, implemented with FFE and auxiliary classifiers, performs better than the eight auxiliary classifiers. This trend is consistent across all metrics and emotions, demonstrating the efficacy of employing an ensemble of expert classifiers that rely on more diversified features. We observe an improvement in the F1-macro metric equal to 4.4% (absolute) compared to the best-performing individual expert classifier (i.e., fear). An additional insight regarding the performance of the individual expert classifiers is that, although inferior to our proposed MoEDE model, their performances are still high when evaluated on the six-class problem in the CREMA-D corpus. This observation suggests that each emotion-based feature extractor is not necessarily biased toward recognizing a single emotion. Instead, they have discriminative information to generalize to multiple emotions after fine-tuning the models for DFER.

The third block of results in Table 5 presents the performance obtained by the baselines. Our first baseline lacks the improvements proposed in this work: emotion-dependent feature extractors and auxiliary classifiers. This baseline, referred to as *single feature extractor* in Table 5, considers as a feature extractor the static classifier trained on the eight classes of the AffectNet corpus, as explained in Section IV. This feature extractor was then fine-tuned for DFER under the same conditions used across all models. Our proposed MoEDE approach, which was implemented with auxiliary classifiers and

TABLE 5. Performance Metrics for DFER on the CREMA-D Corpus. The First Block Presents Results for Our Porposed MoEDE Model With and Without the Auxiliary Classifiers, as Well As, Using the Different SFER Emotion-Dependent Experts (FFT and FTFE, as Noted in Table 3). The Second Block Reports the Individual Performance of the Auxiliary Classifiers Using the Emotion-Dependent Expert Feature Extractors With FFE Models. The Last Block Shows the Performance of Using a Single Feature Extractor (Table 2) and Four Comparable State-of-The-Art Approaches in DFER

	Emotion	F1-macro [% ↑]	F1-weighted [% ↑]	Accuracy [% ↑]	Neutral [% ↑]	Happiness [% ↑]	Sadness [% ↑]	Angry [% ↑]	Fear [% ↑]	Disgust [% ↑]
MoEDE	With auxiliary (FFE) ⋆	**74.5**	**78.9**	**79.3**	**82.3**	**95.6**	**47.6**	**77.5**	67.4	**76.7**
	Without auxiliary (FFE)	72.1	76.8	77.1	80.2	95.0	43.2	73.4	64.7	75.9
	With auxiliary (FTFE) ⋆	73.5	77.9	78.3	81.1	95.3	46.7	76.5	66.8	74.8
	Without auxiliary (FTFE)	71.5	76.3	76.8	80.0	95.3	41.7	72.2	65.1	74.4
Auxiliary classifiers	Neutral	63.9	70.0	71.3	76.3	93.7	27.6	59.1	54.9	71.4
	Happiness	58.9	66.5	68.4	73.0	92.2	10.7	56.1	51.4	70.3
	Sadness	66.7	72.1	72.7	77.1	93.8	35.5	67.1	55.9	70.4
	Anger	64.6	70.7	71.9	74.1	93.4	24.4	69.3	51.4	74.8
	Fear	70.1	74.8	75.1	79.0	94.2	42.5	66.1	64.5	74.2
	Disgust	68.8	74.1	74.8	77.4	93.4	33.3	70.8	62.6	75.2
	Contempt	68.0	72.9	73.4	77.9	92.0	39.5	67.5	59.2	71.9
	Surprise	68.3	73.0	73.6	77.4	91.6	40.7	65.7	58.9	75.1
Baselines	Single feature extractor	70.9	75.2	75.1	79.2	94.1	47.8	73.4	62.3	68.9
	T-LSTM [21]	57.7	64.0	65.0	70.3	91.8	22.3	50.0	46.4	65.4
	AuxFormer [12], [13]	69.3	73.7	73.9	77.2	90.3	42.6	71.2	61.4	73.2
	UAVM [15]	69.2	74.2	74.9	81.0	91.8	42.3	69.0	62.0	69.2
	VAVL [14] ⋆	74.0	78.3	78.7	80.5	95.4	46.1	76.7	**70.1**	75.4

⋆ indicates statistical significance (p-value < 0.05) compared to the single feature extractor model.

an FFE strategy, is significantly better than this baseline across metrics and emotions. The other four baselines correspond to approaches proposed in previous studies: T-LSTM [21], AuxFormer [12], [13], UAVM [15], and VAVL [14] (we describe these methods in Section II). While these models are multimodal, we consider their vision-only implementations. The comparisons are fair, using the five splits of the CREMA-D corpus using the same evaluation metrics. The T-LSTM model [21] trained individual branches for audio and visual modalities, along with a multimodal fusion module, all employing a transformer-based architecture. To enhance the learning of temporal dependencies, an LSTM layer was incorporated before the final linear layer of all blocks. We utilize the visual branch of the pre-trained T-LSTM model to evaluate our testing video sequence. Notably, our best approach outperformed all metrics across all classes, substantially improving 16.8% in F1-macro. The AuxFormer framework [12], [13] allows the evaluation of individual modalities despite being trained based on audio-visual data. The performance of our best MoEDE model outperforms all metrics for all classes, with a notable improvement of 5.2% in F1-macro. For the UAVM model [15], we employ its visual-specific feature extractor followed by its classification layer to evaluate the CREMA-D test set. The performance of the UAVM model is similar to the results observed for the AuxFormer framework. Our best MoEDE approach obtains a 5.3% (absolute) improvement in F1-macro over the UAVM model. The VAVL approach [14] currently has SOTA performance on the CREMA-D. Designed as a versatile audio-visual learning framework capable of accommodating single or multiple modalities, VAVL enabled us to evaluate our best approach using just the visual modality. Our best MoEDE approach closely matches or exceeds the VAVL's performance across all average metrics and five of the six emotional classes, where fear is the exception. These results show the competitive performance achieved by our proposed strategy, indicating the need for feature diversity and generalization.

TABLE 6. Cosine Similarity Between the Last LSTM Step of Different Emotional Feature Extractors ($LSTM_{emo}$)

	Neu	Hap	Sad	Sur	Fea	Dis	Ang	Con
Neu	1.00	0.01	-0.00	0.01	0.01	0.00	0.01	-0.02
Hap	0.01	1.00	-0.02	0.00	-0.00	-0.03	-0.03	0.02
Sad	-0.00	-0.02	1.00	-0.00	0.01	0.02	-0.00	-0.02
Sur	0.01	0.00	-0.00	1.00	-0.02	0.01	-0.01	-0.03
Fea	0.01	-0.00	0.01	-0.02	1.00	-0.03	-0.01	0.01
Dis	0.00	-0.03	0.02	0.01	-0.03	1.00	0.00	-0.01
Ang	0.01	-0.03	-0.00	-0.01	-0.01	0.00	1.00	0.02
Con	-0.02	0.02	-0.02	-0.03	0.01	-0.01	0.02	1.00

Finally, to analyze the diversity of the features, we compute the cosine similarity among the individual expert classifiers by using the 256-dimension vectors extracted after the first LSTM layer of each auxiliary classifier. Table 6 shows the values. The cosine similarities are around 0, showing the separation between the auxiliary classifiers. This finding highlights a significant decrease in similarity among the emotion-dependent classifiers fine-tuned for the DFER task. The average cosine similarity for the static domain reported in Table 4 is around 0.33. This observation highlights the need for more diverse representations to capture the complexity of emotions in the dynamic domain effectively.

C. INDIVIDUAL CONTRIBUTION OF THE EMOTION DEPENDENT EXPERTS

To assess the contribution of each component in our MoEDE architecture, we conduct an ablation analysis on the CREMA-D dataset. Specifically, we evaluated the model's performance by systematically excluding one expert at a time, resulting in eight different configurations, each omitting one of the eight expert classifiers. The performance metrics, presented in

TABLE 7. Performance of MoEDE and Excluded experts/models on Various Metrics for Non-Posed and Spontaneous Expression Recognition

Excluded Expert	F1 macro	F1 weighted	Acc	Hap	Ang	Sad	Neu	Fea	Dis
Neu	74.0	78.3	78.6	95.3	75.9	**47.6**	81.4	**68.0**	76.0
Hap	73.2	77.8	78.2	95.1	76.4	44.6	81.1	66.9	75.3
Sad	73.4	77.9	78.3	**95.9**	76.5	45.3	80.9	65.8	76.0
Sur	74.1	78.6	78.9	95.2	**77.7**	47.2	**82.6**	66.5	75.1
Fear	73.5	77.9	78.3	95.5	77.0	46.3	81.1	64.5	76.5
Dis	73.6	78.2	78.6	95.2	75.9	45.9	81.8	66.9	76.0
Ang	73.5	77.9	78.2	95.4	75.9	47.3	81.2	65.9	75.4
Con	73.9	78.4	78.8	95.9	77.0	47.1	82.2	65.4	76.0
MoEDE	**74.5**	**78.9**	**79.3**	95.6	77.5	**47.6**	82.3	67.4	**76.7**

TABLE 8. Performance Comparison of MoEDE and Other Contemporary Methods on MSP-IMPROV for Non-Posed and Spontaneous Expression Recognition Scenarios

	F1 macro	F1 weighted	Acc	Hap	Ang	Sad	Neu
T-LSTM	61.5	66.9	66.9	77.1	48.5	51.4	68.9
UAVM	56.8	66.7	67.5	79.0	32.9	44.8	70.6
VAVL	58.6	66.3	66.0	79.8	40.9	47.2	66.7
AuxFormer	56.3	60.9	60.9	63.9	49.1	46.3	66.0
SFE	50.8	62.7	64.4	75.8	13.7	46.3	67.6
MoEDE	57.2	67.9	69.3	83.2	30.3	45.0	70.3

Table 7, are based on the mean results from five independent runs.

Table 7 shows that the MoEDE model consistently achieves the highest overall accuracy and F1-scores. However, in some cases, certain configurations—where one expert is excluded—surpass the complete model's performance in specific emotions. For example, configurations lacking one expert show improvements of 0.3% in happiness, 0.6% in fear, and 0.2% in anger. These variations indicate that, while each expert significantly contributes to the model's overall performance, the exclusion of any one expert often results in a slight drop in the model's metrics. This result emphasizes the importance of each expert in our architecture, with each component playing a critical role in optimizing performance.

D. GENERALIZATION TO OTHER DATASETS

Our MoEDE model leverages feature extractors trained on the AffectNet dataset, which comprises a wide range of 'in-the-wild' expressions. This design enables the model to generalize effectively to non-posed and spontaneous expression recognition scenarios. To validate this capability, we evaluate MoEDE on the MSP-IMPROV dataset [2], which captures emotional behaviors during spontaneous dyadic improvisations. For consistency, we use the same hyperparameters applied in our CREMA-D experiments.

Table 8 presents the performance of MoEDE compared to other popular methods and our *single feature extractor* (SFE) baseline on the MSP-IMPROV corpus. The reported values represent the mean results across six independent runs, each using subject-independent splits. MoEDE achieves the highest accuracy at 69.3% , outperforming the next best model, UAVM, by 1.8% . It also attains the highest score on F1-weighted metrics.

While MoEDE's F1-macro score is lower than the results obtained by the T-LSTM and VAVL models, these models exhibit approximately 3% lower accuracy. We hypothesize that fine-tuning hyperparameters specifically for the MSP-IMPROV development set could further improve MoEDE's performance, a strategy likely employed by the other models.

VI. CONCLUSION

This paper presented our proposed MoEDE method that leverages emotion-dependent feature extractions to enhance the diversity of facial features in DFER. Our findings indicate the potential of fine-tuned, emotion-dependent models to improve the classification of facial expressions in videos across diverse emotional states. These results underscore the importance of model architecture choices and training strategies in the field of emotion recognition. The results show the effectiveness of our approach on the CREMA-D datasets, showing an F1-macro improvement between 0.5% and 16.8% (absolute) when compared to SOTA approaches. Additionally, we showed that by fine-tuning and improving the feature extractor in the static domain, the F1-macro of our proposed DFER approach drops by 1% . This result highlights that improved features designed to improve SFER, do not necessarily improve DFER models.

For future work direction, we plan to explore different approaches to integrating our DFER into multimodal audio-visual systems. Our results demonstrate that our approach can rival the performance of multimodal models when evaluated on a single modality. Therefore, we hypothesize that incorporating our contributions into multimodal systems for FER can result in even more effective systems. Additionally, we want to explore using video or image-based gating or routing mechanisms to reduce the computational cost of our proposed approach by using a subset of the expert models at the frame or video level. Finally, although the MoEDE model is primarily designed for recognizing emotional categories, we believe it also holds potential in the recognition of other emotion-related tasks such as detecting micro-expression.

REFERENCES

[1] T. Agustin, M. H. Purwidiantoro, and M. L. Rahmadi, "Enhancing facial expression recognition through ensemble deep learning," in *Proc. IEEE 5th Int. Conf. Cybern. Intell. System*, Pangkalpinang, Indonesia, 2023, pp. 1–6.

[2] C. Busso et al., "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, Jan.–Mar. 2017.

[3] C. Busso and S. Narayanan, "Interplay between linguistic and affective goals in facial expression during emotional utterances," in *Proc. 7th Int. Seminar Speech Prod.*, Ubatuba-SP, Brazil, Dec. 2006, pp. 549–556.

[4] C. Busso and S. Narayanan, "Interrelation between speech and facial gestures in emotional utterances: A single subject study," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2331–2347, Nov. 2007.

[5] H. Cao et al., "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, Oct.–Dec. 2014.

[6] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial expression recognition in video with multiple feature fusion," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 38–50, Jan.–Mar. 2018.
[7] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, "Deep learning-based facial emotion recognition for human–computer interaction applications," *Neural Comput. Appl.*, vol. 35, no. 32, pp. 23311–23328, 2023.
[8] D. W. Cunningham and C. Wallraven, "Dynamic information for the recognition of conversational expressions," *J. Vis.*, vol. 9, no. 13, pp. 1–17, Dec. 2009.
[9] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 248–255.
[10] L.-N. Do, H.-J. Yang, H.-D. Nguyen, S.-H. Kim, G.-S. Lee, and I.-S. Na, "Deep neural network-based fusion model for emotion recognition using visual data," *J. Supercomputing*, vol. 77, pp. 1–18, 2021.
[11] T. Dozat, "Incorporating Nesterov momentum into Adam," in *Proc. Workshop Track Int. Conf. Learn. Representations*,San Juan, Puerto Rico, May 2015, pp. 1–4.
[12] L. Goncalves and C. Busso, "AuxFormer: Robust approach to audiovisual emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Singapore, May 2022, pp. 7357–7361.
[13] L. Goncalves and C. Busso, "Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2156–2170, Oct.–Dec. 2022.
[14] L. Goncalves, S.-G. Leem, W.-C. Lin, B. Sisman, and C. Busso, "Versatile audiovisual learning for handling single and multi modalities in emotion regression and classification tasks," *IEEE Trans. Affect. Comput.*, to be published.
[15] Y. Gong et al., "UAVM: Towards unifying audio and visual models," *IEEE Signal Process. Lett.*, vol. 29, pp. 2437–2441, 2022.
[16] A. Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech 2020*, Shanghai, China, Oct. 2020, pp. 5036–5040.
[17] M. Hao, W.-H. Cao, Z.-T. Liu, M. Wu, and P. Xiao, "Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features," *Neurocomputing*, vol. 391, pp. 42–51, 2020.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun./Jul. 2016, pp. 770–778.
[19] A. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," Apr. 2017, *arXiv:1704.04861*.
[20] E. J. Hu et al., "Lora: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Representations*, 2022.
[21] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," in *Proc. 2020 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 3507–3511.
[22] C. Jia, C. L. Li, and Z. Ying, "Facial expression recognition based on the ensemble learning of CNNs," in *Proc. 2020 IEEE Int. Conf. Signal Process. Commun. Comput.*, Macau, China, 2020, pp. 1–5.
[23] Y. Kim and E. M. Provost, "ISLA: Temporal segmentation and labeling for audio-visual emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 196–208, Apr.–Jun. 2019.
[24] D. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Jul. 2009.
[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, CA, USA, Dec. 2012, vol. 25, pp. 1097–1105.
[26] J. Li, J. Nie, D. Guo, R. Hong, and M. Wang, "Emotion separation and recognition from a facial expression by generating the poker face with vision transformers," *IEEE Trans. Comput. Social Syst.*, pp. 1–15, 2024.
[27] S. Li, W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3.
[28] J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, and A. Huang, "Poster: A simpler and stronger facial expression recognition network, *Pattern Recognit.*, vol. 157, 2024.
[29] S. Mariooryad and C. Busso, "Factorizing speaker, lexical and emotional variabilities observed in facial expressions," in *Proc. IEEE Int. Conf. Image Process.*, Orlando, FL, USA, Sep./Oct. 2012, pp. 2605–2608.
[30] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. 2016 IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–10.
[31] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan.–Mar. 2019.
[32] A. Salman and C. Busso, "Dynamic versus static facial expressions in the presence of speech," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Buenos Aires, Argentina, May 2020, pp. 436–443.
[33] A. N. Salman and C. Busso, "Style extractor for facial expression recognition in the presence of speech," in *Proc. IEEE Int. Conf. Image Process.*, Abu Dhabi, UAE, Oct. 2020, pp. 1806–1810.
[34] M. Sandler et al., "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 4510–4520.
[35] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2132–2143, Oct.–Dec. 2022.
[36] H. Siqueira, S. Magg, and S. Wermter, "Efficient facial feature learning with wide ensemble-based convolutional neural networks," in *Proc. AAAI Conf. Artif. Intell.*, New York, NY, USA, 2020, vol. 34, pp. 5800–5809.
[37] L. Sun et al., "Svfap: Self-supervised video facial affect perceiver," in *Proc. Int. Conf. Multimedia*, 2023.
[38] L. Sun, Z. Lian, B. Liu, and J. Tao, "MAE-DFER: Efficient masked autoencoder for self-supervised dynamic facial expression recognition," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 6110–6121.
[39] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Mach. Learn. Res.*, Jun. 2019, vol. 97, pp. 6105–6114.
[40] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, pp. 6105–6114, Jun. 9–15, 2019.
[41] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.
[42] S. Zhang and Z. Others, "A dual-direction attention mixed feature network for facial expression recognition," *Electronics*, vol. 12, no. 17, 2023, Art. no. 3595.
[43] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. CVPR*, Jun. 2016.