Efficient AudioVisual Fusion Architectures for Emotion Recognition

Leila Ben Letaifa, Amine Bohi

▶ **To cite this version:**

Leila Ben Letaifa, Amine Bohi. Efficient AudioVisual Fusion Architectures for Emotion Recognition. Neural Computing and Applications, inPress, ⟨10.1007/s00521-026-12056-5⟩. ⟨hal-05553820⟩

HAL Id: hal-05553820

https://hal.science/hal-05553820v1

Submitted on 16 Mar 2026

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright - All rights reserved

Efficient AudioVisual Fusion Architectures for Emotion Recognition

Leila Ben Letaifa[1*] and Amine Bohi[2†]

[1]CESI LINEACT, UR 7527, Vandoeuvre Les Nancy, 54500, France.
[2]CESI LINEACT, UR 7527 , Dijon, 21800, France.

*Corresponding author(s). E-mail(s): lbenletaifa@cesi.fr;
Contributing authors: abohi@cesi.fr;
†The authors contributed equally to this work.

Abstract

Emotion recognition plays a critical role in the development of adaptive, human-aware intelligent systems. In this work, we propose an end-to-end audiovisual emotion recognition framework that integrates speech signals and facial expressions using lightweight deep learning architectures. To develop the end-to-end architecture, we first benchmark several pretrained convolutional neural networks, employing confidence interval estimation to statistically evaluate the trade-off between recognition accuracy and model complexity. EfficientNetV2-B0 is identified as the most effective backbone for facial emotion recognition and is subsequently adopted as the feature extractor in our audiovisual framework. Achieving a lightweight and efficient audiovisual emotion recognition system requires optimizing accuracy, robustness, and model size. We address this by proposing three progressively refined architectures that combine model-based and late fusion techniques. The baseline model employs a transformer-based architecture for audiovisual fusion. To handle potential modality absence, we introduce a variant that enhances the modeling of modality-specific characteristics. This is further strengthened through the integration of self-attention mechanisms within each modality, enabling the system to capture both cross-modal correlations and intra-modal dynamics effectively. Extensive experiments conducted on the RAVDESS dataset demonstrate that our proposed architectures outperform existing state-of-the-art methods on this benchmark. Furthermore, the models exhibit strong performance with a low memory footprint, making them well-suited for resource-constrained devices.

Keywords: Audiovisual Emotion Recognition, Lightweight Neural Networks, Pretrained models, Fusion Techniques, Attention Mechanisms, Efficient Deep Learning, Facial Expression Recognition

1 Introduction

Emotion recognition has attracted increasing interest in artificial intelligence (AI) research, with a wide range of application scenarios, from emotion-aware chatbots and consumer behavior analysis to assistive humanoid robots [1–3]. This task involves inferring a person's emotional state based on speech, appearance, behavior, and other indicators. When considered individually, modalities such as text, audio, or video often provide limited or ambiguous information. However, significant improvements can be achieved through multimodal fusion, which leverages the complementary strengths of each modality. For instance, in noisy environments, facial expressions may serve as more reliable indicators of emotion than speech; conversely, when a speaker's face is partially obscured or turned away, vocal cues can offer greater insight. As a result, the use of multimodal systems has become increasingly important for improving the reliability of emotion recognition [1, 4, 5].

Conventional pattern recognition systems typically follow a two-stage pipeline: feature extraction from input data, followed by classification. These features are traditionally handcrafted, making them highly dependent on data characteristics and requiring extensive domain-specific expertise. In the domain of facial expression recognition (FER), classical methods are generally divided into two categories: geometric-based approaches, which analyze the shape and spatial configuration of facial landmarks during expressions [6, 7]; and appearance-based techniques, which capture local or global texture variations such as wrinkles and furrows [8, 9]. Likewise, speech emotion recognition (SER) often relies on low-level handcrafted descriptors, including pitch, energy, formants, Mel-Frequency Cepstral Coefficients (MFCCs), and prosodic features [10–12]. These features are commonly fed into traditional classifiers such as Support Vector Machines (SVM), k-Nearest Neighbors (kNN), or Multilayer Perceptrons (MLP). Decoupling feature extraction from classification limits the system's ability to optimize performance globally—particularly when dealing with complex, heterogeneous, or multimodal data. Moreover, it relies on specialized human expertise, which is often limited and not easily accessible. To overcome these challenges, end-to-end deep learning approaches have emerged as a compelling alternative. By learning representations directly from raw inputs, deep neural networks unify feature extraction and classification into a single trainable framework. This integrated paradigm offers greater flexibility, reduces reliance on manual feature design, and enhances generalization across diverse modalities and application scenarios. Consequently, end-to-end deep learning has achieved remarkable successes across various pattern recognition tasks, including emotion recognition [13–18].

Recently, research on end-to-end systems has increasingly examined the use of transfer learning to extract high-level features from pretrained deep neural networks. These neural network models are often trained on large-scale datasets and offer strong and reliable feature extraction capabilities. Nevertheless, their considerable computational demands and large parameter sizes present major obstacles to deployment on resource-limited mobile platforms [19–21].

Building on these advances, in this study we explore the use of pretrained neural networks for feature extraction in FER, with a focus on balancing classification performance and model complexity. Multiple FER systems, each based on a different

pretrained architecture, were evaluated under consistent experimental conditions. A benchmarking analysis with statistical confidence intervals identified the most effective model, which was then integrated as the visual feature extractor in a multimodal Audiovisual Emotion Recognition (AVER) framework.

Designing a lightweight end-to-end AVER system presents significant challenges, particularly in balancing performance, robustness, and computational efficiency. An optimal architecture must:

- jointly process both visual and auditory modalities,
- remain robust if one modality is missing,
- integrate pretrained feature extractors seamlessly,
- and remain compact enough for deployment on resource-constrained mobile platforms.

To address all these requirements, we propose novel end-to-end AVER architectures that leverage both intermediate and late fusion strategies. Intermediate fusion is performed using a Transformer block with a multimodal attention mechanism, enabling efficient integration of audiovisual cues. To enhance robustness, unimodal features are combined using a late fusion approach before classification. Temporal dependencies within each modality are captured via self-attention mechanisms, whose outputs are similarly merged prior to the final decision stage. Model size was carefully controlled during design to guarantee compatibility with deployment on autonomous mobile platforms. The proposed architectures were implemented and evaluated on the RAVDESS audiovisual dataset, with experimental results demonstrating superior performance over existing methods.

The paper is organized as follows: the first section reviews the state of the art in emotion recognition systems, with particular attention given to FER and AVER. The following section outlines the principles and methodology underlying the conception and development of our deep learning model architectures. Lastly, we present the experimental protocol and provide a detailed analysis and comparison of the obtained results.

2 Related Work

In the existing literature, numerous studies on emotion recognition have been carried out using both unimodal and multimodal approaches. In this section, we present the state of the art in facial expression recognition (FER) as well as in audiovisual emotion recognition (AVER).

2.1 Facial Expression Recognition

Facial expression recognition (FER) has witnessed remarkable progress with the emergence of deep learning. Convolutional Neural Networks (CNNs), attention mechanisms, and transformer-based architectures have enabled substantial performance improvements on benchmarks such as FER2013 [22], RAF-DB [23], and AffectNet [24]. Various strategies have been proposed, including attention-guided learning [25, 26], multi-branch architectures [27], adaptive loss functions [28], and dual-attention fusion

techniques [29]. Transformer-based models like POSTER [30] and HLA-ViT [31] further enhanced robustness to pose and occlusion. In [17, 18], the authors introduced the EmoNeXt architecture which integrates Spatial Transformer Networks for alignment correction, Squeeze-and-Excitation (SE) blocks for channel-wise feature recalibration, and self-attention regularization. This hybrid design enables precise localization of expressive regions and effective feature modulation, achieving state-of-the-art results across FER2013, CK+, and AffectNet.

Despite the growing accuracy of FER systems, many rely on large models with high memory and computational costs, which restricts their deployment in real-time and embedded contexts, such as human-robot interaction. To address this, lightweight architectures like MobileNet, EfficientNet, and ConvNeXt-Tiny , as well as transfer learning techniques have been explored [32–34]. Having established the visual backbone, we now turn to the multimodal fusion strategies that form the core of our audiovisual architectures.

2.2 Audiovisual Emotion Recognition

While unimodal approaches, such as those based solely on facial expressions, have shown promising results, they often struggle to capture the full complexity and subtlety of human emotions—particularly in spontaneous, real-world scenarios. To address these limitations, the research community has increasingly embraced multimodal strategies, combining complementary cues from different sources—especially audio and visual signals—to enhance the robustness and generalizability of emotion recognition systems. The literature is rich in studies demonstrating the benefits of multimodal emotion recognition compared to unimodal approaches. In the context of audiovisual systems, Xiang et al. [35] reported unimodal evaluations on the RAVDESS dataset that exhibit inferior performance compared to multimodal fusion strategies. Similar conclusions were drawn by Tzirakis et al. [36], who showed on the RECOLA dataset that audio and visual modalities convey complementary emotional information, and that their joint modeling consistently outperforms unimodal systems. More recently, Palmero et al. [1] reported comparable observations in the context of human–virtual agent interaction, demonstrating that the fusion of speech and facial expressions leads to improved emotion recognition performance over audio-only or visual-only approaches. In support of this trend, numerous multimodal datasets have emerged, such as IEMOCAP [37], Aff-Wild2 [38], RAVDESS [39], CREMA [40], EMPATHIC corpus [41, 42] and CG-MER dataset [43][44].

One early exploration of audiovisual fusion was conducted by Lian et al. [45], who proposed combining handcrafted and deep features from audio, video, and text using both temporal and non-temporal classifiers. A Beam Search Fusion (BS-Fusion) strategy was employed to optimize modality integration. Their system achieved 60.34% accuracy on the EmotiW 2018 test set—only 1.5% below the top-ranked entry—demonstrating the potential of hybrid and ensemble techniques in multimodal affect analysis. Zhang et al. [46] introduced a spatiotemporal network combining a 2D-CNN with a Temporal Convolutional Network (TCN) for the visual stream and parallel TCNs for the audio stream. A leader-follower attention mechanism was used to prioritize more reliable modalities. The system, evaluated on the Aff-Wild2 dataset

using Concordance Correlation Coefficient (CCC), showed clear performance gains in continuous valence and arousal estimation tasks. In a similar vein, Lee et al. [47] extended the BERT architecture to multimodal inputs by integrating audio and visual features via attention-based modules. Their approach, tested on CMU-MOSI, CMU-MOSEI, and IEMOCAP, achieved state-of-the-art results, underlining the effectiveness of attention-driven heterogeneous feature fusion for fine-grained emotion and sentiment analysis. Luna et al. [48] presented a bimodal system based on speech and facial expressions. For speech, they fine-tuned a pre-trained XLSR-Wav2Vec2.0 transformer with a multilayer perceptron head, outperforming simple embedding extraction. For the facial stream, Action Units were extracted and processed using sequential models, which slightly outperformed static ones. Late fusion of both modalities yielded 86.70% accuracy on the RAVDESS dataset with subject-wise 5-fold cross-validation. The authors also emphasized the potential benefits of focusing on high-emotional-load frames in future visual models. A related contribution by the same group [39] proposed another multimodal AVER framework, this time using CNN-14 from the PANNs framework for speech and a pre-trained Spatial Transformer Network (STN) followed by a bi-LSTM with attention for facial analysis. Again, fine-tuning was found superior to embedding extraction. Their late fusion approach achieved 80.08% accuracy on RAVDESS under the same evaluation protocol, and highlighted the limitations of applying frame-based architectures directly to video-based classification. Gonçalves et al. [49] focused on robustness under adverse conditions such as noise and occlusion. Their method relied on auxiliary networks, transformers, and a curriculum-based training strategy to enhance modality alignment and temporal modeling. The system demonstrated improved resilience in challenging environments. More recently, Xiang et al. [35] introduced the MultiMAE-DER framework, a masked autoencoding approach that models spatio-temporal correlations in audiovisual emotion recognition. The model was fine-tuned using six fusion strategies and evaluated on RAVDESS, CREMA-D, and IEMOCAP. It outperformed several supervised and self-supervised baselines, with up to +4.41% improvement in weighted average recall on RAVDESS. These results confirm the benefit of leveraging self-supervised learning and optimized fusion in dynamic, multimodal settings.

These recent efforts illustrate the growing importance of attention mechanisms, temporal modeling, and robust fusion strategies in audiovisual emotion recognition. They also highlight the growing trend of leveraging pre-trained architectures and attention-based strategies to advance the field. In this context, we propose a novel audiovisual emotion recognition approach that builds on these principles, designed to efficiently integrate facial and vocal modalities in a unified learning framework.

3 Methodology

This section outlines the methodology used to develop our emotion recognition systems. We begin with an efficient FER backbone built on pretrained CNNs, serving as a feature extractor. This unimodal benchmark helps identify the most effective pre-trained model. Building on this, we explore various fusion strategies to integrate speech and facial cues into a lightweight, robust, and accurate system. Specifically, we propose

three audiovisual architectures: a baseline cross-attention model, a hybrid model with modality-specific concatenation, and an advanced design leveraging self-attention. These progressively refined architectures aim to balance performance, robustness, and computational efficiency for real-world applications.

3.1 Facial Emotion Recognition Backbone

Pretrained CNNs have shown remarkable performance in facial emotion recognition (FER), particularly when used with transfer learning [17, 18]. These models, trained on large-scale datasets such as ImageNet [50], can effectively capture visual features that are generalizable across different domains.

As shown in Fig. 1, our FER system relies on such pretrained CNNs to extract spatial features from facial images. The approach involves discarding the original classification layers of the network and retaining only the convolutional feature extractor. These convolutional layers are capable of detecting low-level visual structures such as edges, shapes, and textures that are crucial for recognizing facial expressions. The feature maps are flattened and fed into a lightweight classification network composed of fully connected layers, whose final layer predicts several emotion classes such as anger, disgust, fear, happiness, sadness, surprise, and neutrality.

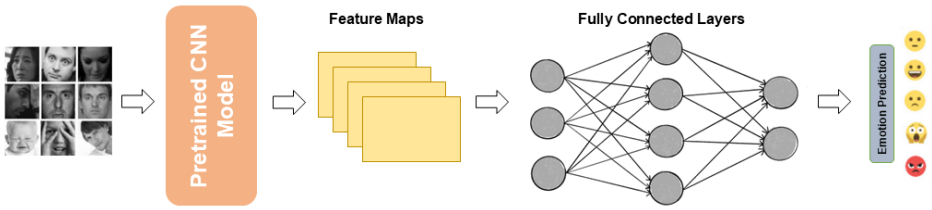

Fig. 1 Overview of a FER system architecture using transfer learning.

Deploying the FER system on mobile devices, which are typically constrained in computational resources, necessitates the use of a compact model. In this context, the overall model size is determined by the total number of parameters, including those of the pre-trained feature extractor model and the fully connected network classifier. This system forms the visual processing backbone of the unimodal FER pipeline. The same pretrained model is further utilized for visual feature extraction within the end-to-end AVER architecture proposed in this study.

3.2 Multimodal fusion techniques

In multimodal emotion recognition, three primary fusion strategies are commonly employed to integrate information from different modalities: early fusion, intermediate fusion, and late fusion [4, 51]. Each approach offers specific advantages and limitations, depending on how the data is combined within the model. The operational principles of the three modalities are depicted in Fig. 2.

- **Early Fusion**: Also known as feature-level fusion, this method combines information from multiple modalities at an early stage of the processing pipeline [5]. Features derived from audio, visual, or other signals are typically merged using operations such as concatenation or addition before being passed to the learning model. Despite its ability to enable joint representation learning, early fusion faces difficulties in capturing complex inter-modal dependencies. It also demands precise temporal alignment across modalities and is particularly sensitive to noise. In deep learning-based AVER systems, the overall model size under early fusion schemes mainly depends on the complexity of the downstream classification architecture.
- **Late Fusion**: Known as decision-level fusion, this method combines the outputs of separate models, each trained on an individual modality. Each modality independently generates a prediction, and these outputs are then aggregated to produce the final decision. Late fusion is generally more robust to variations across modalities and can effectively handle missing or degraded data [5]. However, it may fail to capture complex interactions between modalities—since they are modeled independently—potentially limiting the model's ability to exploit synergistic relationships. With respect to system size, it is often substantial, as it is closely tied to the cumulative sizes of the models associated with each modality.
- **Intermediate Fusion**: Commonly known as representation-level or model-level fusion, this strategy combines modality-specific features at an intermediate stage within the neural network architecture. Each modality is first processed independently to extract its own features, which are then combined within the model. This technique maintains the unique characteristics of each modality while leveraging the complementary information they provide, offering a balanced trade-off between early and late fusion [1, 4]. Moreover, this approach performs emotion classification with a single unified model. For all these reasons, it was adopted for the baseline AVER architecture in this study.

3.3 Audiovisual model architectures

Model-level fusion, compared to feature- and decision-level fusion, captures multimodal interactions within the model and more effectively harnesses the power of deep neural networks. We propose a baseline architecture that uses a Transformer model to fuse audio-visual modalities at the model level. Multi-head attention produces multimodal intermediate emotional representations from a shared semantic feature space after encoding the audio and visual inputs. To increase the system's robustness—when a modality is missing, for example—a late fusion mechanism is integrated into the architecture, using either the features of each modality or the self-attention mechanism.

3.3.1 The Transformer model

The transformer model [52] is a sequence-to-sequence architecture designed to map an input sequence $(x_1, x_2, ..x_T)$ to an output sequence $(y_1, y_2, .., y_L)$. Its structure consists of two main components: the encoder and the decoder. The encoder takes the input sequence and transforms it into an intermediate sequence of encoded features

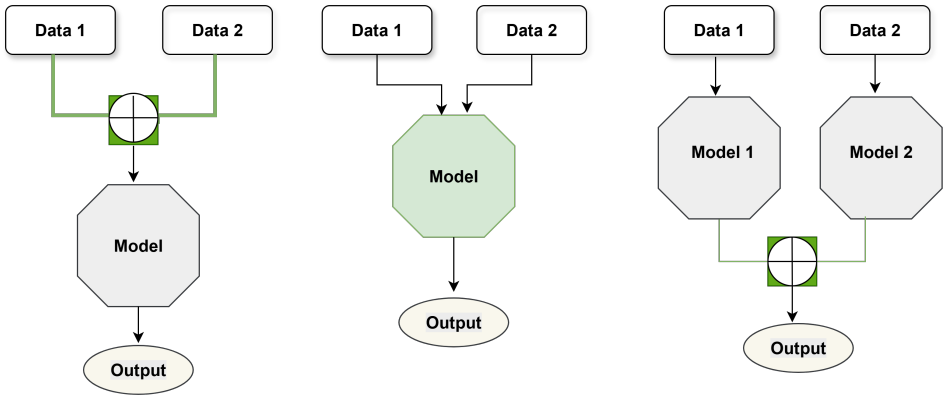

Fig. 2 Multimodal fusion techniques : early fusion (left), late fusion (right) and intermediate fusion (middle). Here, Data 1 and Data 2 represent the input modalities, Model denotes the classification model, and Output corresponds to the system's classification scores.

$(h_1, h_2, .., h_N)$. On the other hand, the decoder generates predictions for each output y_l based on the encoded features $(h_1, h_2, .., h_N)$ and the previously decoded outputs $(y_1, y_2, .., y_{l-1})$. Both the encoder and the decoder comprise a stack of multi-head attention (MHA) and feedforward network (FFN) layers, allowing the model to capture contextual dependencies and make accurate predictions [53]. Each layer is accompanied by a residual connection and normalization. Fig. 3 provides an illustration of the encoder block structure.

The layers of the encoder refine the representation of the input sequence with a suite of multi-head, self-attention, and linear transformations [53]. The self-attention operation allows frames to gather context from all timesteps and build an informative sequence at a high level. Specifically, the inputs of each layer are projected into queries Q, keys K, and values V with $Q \in \mathbb{R}^{t_q * d_q}$, $K \in \mathbb{R}^{t_k * d_k}$, and $V \in \mathbb{R}^{t_v * d_v}$. $t*$ denotes the number of elements in the inputs, and $d*$ the corresponding dimensions. Usually, these are $t_k = t_v$ and $d_q = d_k$.

Scaled Dot-Product Attention [52] is then computed as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

The MHA is applied to take advantage of the different representations that are simultaneously present. The MHA is obtained by performing this calculation h times. h is the number of heads.

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W_0 \tag{2}$$

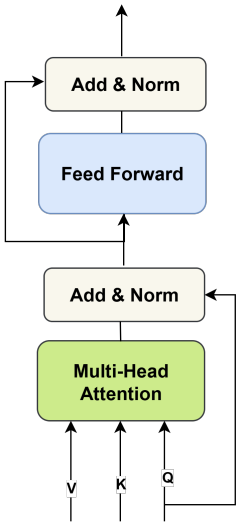

Fig. 3 Transformer model encoder block.

where

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

The projection matrices are $W_i^Q \in \mathbb{R}^{d_{model}*d_q}$, $W_i^K \in \mathbb{R}^{d_{model}*d_k}$, $W_i^V \in \mathbb{R}^{d_{model}*d_v}$, and $W^0 \in \mathbb{R}^{h*d_v*d_{model}}$. In this work, $d_k = d_q = d_v = d_{model}/h$.

The outputs of the MHA go through a two-layer position-wise feedforward network (FFN) with hidden size d_{ff} [52].

$$FFN(x) = W_2 ReLU(W_1 x + b_1) + b_2 \tag{4}$$

$b_1 \in \mathbb{R}^{d_{ff}}$ and $b_2 \in \mathbb{R}^{d_{model}}$ are the biases. The weight matrices are $W_1 \in \mathbb{R}^{d_{model}*d_{ff}}$ and $W_2 \in \mathbb{R}^{d_{ff}*d_{model}}$.

3.3.2 Baseline architecture

The baseline architecture leverages a pretrained model for visual feature extraction, while audio inputs are represented using Mel-Frequency Cepstral Coefficients (MFCCs). A pretrained model is employed exclusively for the visual stream, serving as an initial step in validating the proposed AVER systems.

As outlined in Section 3.2, a model-based fusion strategy was selected for integrating speech and facial modalities due to its advantages over early and late fusion approaches. Multimodal fusion is implemented via a Transformer-based architecture utilizing MHA mechanisms. Specifically, two cross-attention blocks are introduced:

- an Audio-Visual (AV) block, where visual features attend to audio emotion cues,
- and a Visual-Audio (VA) block, where audio features attend to visual embeddings.

Following attention-based fusion, pooled feature representations from both modalities are concatenated and fed into the classification layer. An overview of the full pipeline is illustrated in Fig. 4.

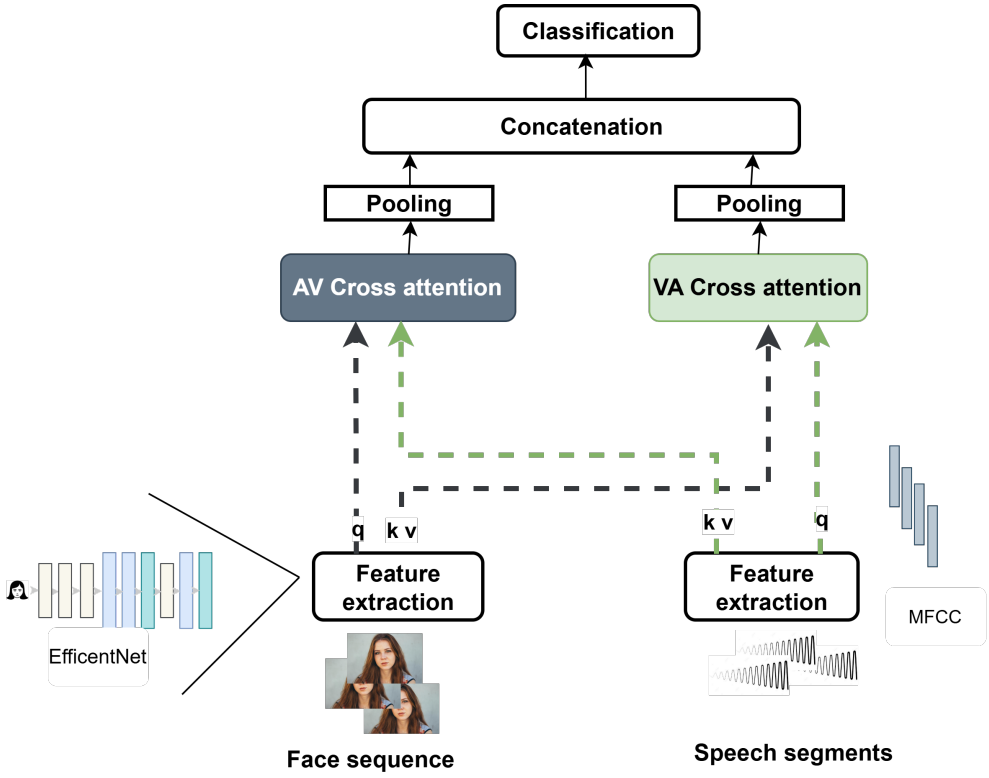

Fig. 4 Baseline audiovisual architecture composed of three main stages: feature extraction, cross-modal Transformer-based fusion (AV and VA), and classification. The audio stream is represented by a green dotted line, while the visual stream is shown as a black dotted line. Both modalities are fused within the Transformer block.

The size of the proposed model is primarily determined by two key components. The first is the pretrained backbone, which accounts for the majority of the parameters, as it is responsible for extracting high-level feature representations from the visual input. The second component comprises the Transformer-based AV and VA cross-attention blocks, which introduce additional parameters to capture fine-grained interactions between the audio and visual streams. Together, these components constitute the bulk of the model's parameters and largely define its overall size.

3.3.3 Input data incorporation

Although the cross-modal attention blocks are effective at capturing jointly complex audio and video cues, relying solely on them is not always sufficient to ensure the

robustness of the model, particularly in scenarios where one modality may be missing or degraded. To overcome this limitation, we integrated two additional monomodal blocks—one for each modality—originating from the feature extraction stage. These blocks are designed to reinforce the features of the available modality, ensuring that essential information is preserved even when cross-modal correlations are weak or absent.

The outputs of the monomodal blocks are concatenated with those of the cross-attention Transformer blocks, enabling the model to preserve and leverage modality-specific information in addition to inter-modal correlations. This mechanism resembles a late fusion strategy, where the final prediction considers the contributions of each modality separately. Fig. 5 shows the architecture of the corresponding system. Integrating features from both modalities has a negligible impact on model size, as it involves only the concatenation of output vectors.

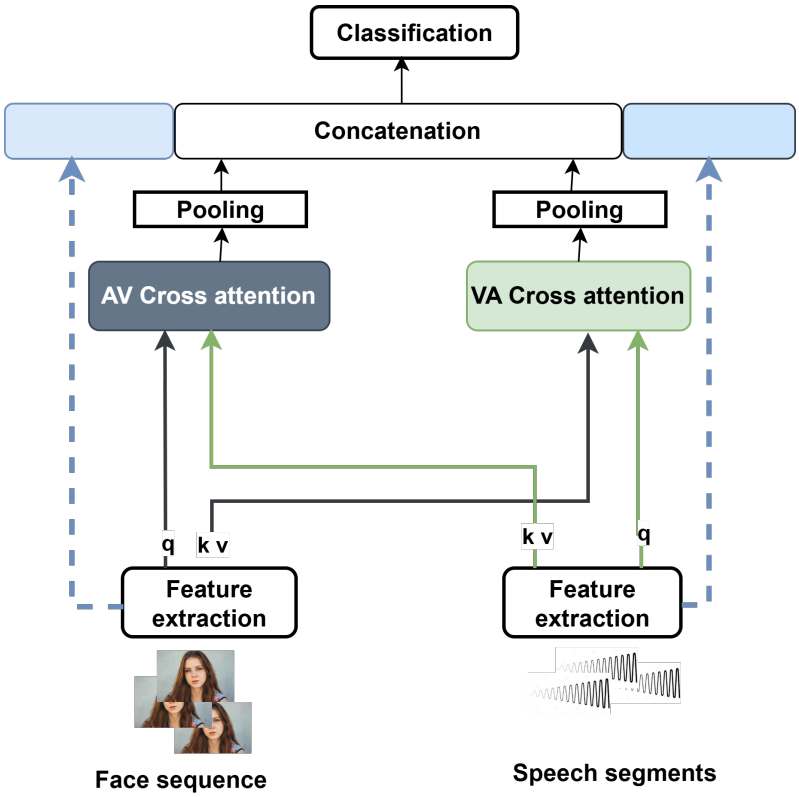

Fig. 5 Input Data Incorporation architecture is an extension of the baseline model and relies on feature concatenation to integrate monomodal representations. For each modality, the output of the feature extraction stage (represented by a dotted blue line) is concatenated with the outputs of the Transformer blocks before being fed into the classification module.

3.3.4 Self-attention information

In this subsection, we enhance the model architecture by integrating two self-attention blocks, as illustrated in Fig. 6. These blocks are independently applied to each modality prior to the final classification stage, consistent with the previous architecture. Their primary function is to capture temporal dependencies specific to each data stream, independently from the inter-modal correlations learned by the Transformer blocks.

This integration can be regarded as a form of late fusion, which improves the model's robustness in scenarios where modalities are weakly correlated or when one modality is unavailable.

Collectively, this hybrid architecture, combining intra-modal self-attention with inter-modal attention mechanisms, enhances overall performance by fully leveraging both modality-specific information and their cross-modal interactions.

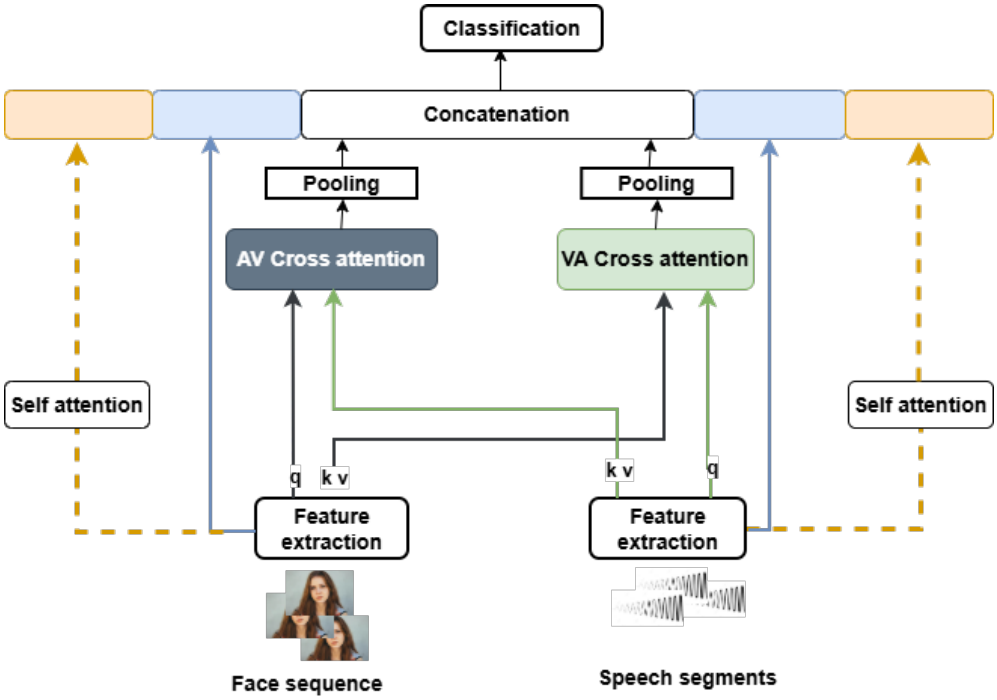

Fig. 6 Self-Attention Information architecture is an extension of the two previous architectures and is based on self-attention for integrating unimodal features. For each modality, the output of the self-attention block is concatenated with the outputs of the Transformer blocks. This process is highlighted in orange in the figure.

3.3.5 Architectural Overview

The main characteristics of the three proposed architectures are summarized in Table 1. The baseline architecture employs a cross-modality attention (CA) mechanism to integrate audio and video modalities within the model.
The second architecture, denoted as Input Data Incorporation (ID), extends the baseline by introducing a concatenation layer aimed at improving robustness in the presence of missing data or weak inter-modality correlations. Specifically, the audio stream and the image stream are concatenated with the transformer block outputs prior to being fed into the classification layer.
The third architecture further extends the second by applying a self-attention (SA) mechanism independently to each modality before concatenation. This approach also implements a late fusion strategy, allowing the model to capture modality-specific temporal dynamics prior to multimodal integration.

Architecture	**Layer**	**Role**	**Contribution**
Baseline	CA	Audio–visual fusion	Models inter-modal correlations through explicit audio–visual interactions.
Input Data Incorporation	ID	Modality-specific information integration	Late-fusion strategy improving robustness when modalities are weakly correlated or partially missing.
Self-Attention Information	SA	Temporal modeling within each modality	Captures intra-modal temporal dependencies while preserving complementary modality information.

Table 1 Comparison of the roles and contributions of the three proposed architectures. CA, ID, and SA denote Cross-Attention, Input Data Incorporation, and Self-Attention, respectively.

4 Experiments and results

This section presents the experimental setup and results supporting the development of our facial and audiovisual emotion recognition systems. The overarching goal is to design an accurate, lightweight, and robust end-to-end AVER framework. To achieve this, the following steps are undertaken: (i) selecting a pretrained visual feature extractor that offers a balance between accuracy and model compactness; (ii) developing an efficient and low-overhead AVER system; and (iii) enhancing its robustness in the presence of missing or incomplete modality data.

4.1 Pretrained model selection

This section builds upon our previous work [33], where we benchmarked various deep learning models for facial expression recognition using the FER2013 dataset [22]. The

goal of those experiments was to identify the most suitable model for real-time deployment in a human–machine interface and on the humanoid robot Tiago++[1], which is limited to a maximum model size of 150 MB. The statistical analysis conducted in this section, including confidence interval estimation, allows us to compare model performance with greater robustness.

4.1.1 Face emotion recognition

We evaluated the effectiveness of multiple pretrained convolutional neural network (CNN) models available in the Keras library[2], all initially trained on the ImageNet dataset [50], and are widely adopted for various visual classification tasks. As detailed in subsection 3.1, we adopted a transfer learning approach by discarding the original classification head of each model and fine-tuning the remaining convolutional layers on the FER2013 dataset. The set of evaluated models includes MobileNet [54], DenseNet201 [55], ResNet152V2 and ResNet101 [56, 57], Xception [58], EfficientNetV2-B0 [59], InceptionV3 and InceptionResNetV2 [60], VGG16 and VGG19 [61], as well as several variants of ConvNeXt from Tiny to XLarge [62]. Each model was followed by a fully connected neural network (FC) consisting of an input layer with 5,120 units, a hidden layer of 768 neurons, and an output layer with 8 units corresponding to the target classes. This FC network comprises approximately 31.45 million parameters.

To enhance generalization and robustness, all models were trained using an extensive data augmentation pipeline including geometric transformations (rotation, shift, zoom, horizontal flipping), brightness/contrast adjustments, and Random Erasing to simulate occlusions. Additionally, variations in resizing and recropping were employed to increase tolerance to face positioning variance. Optimization was performed using the Adam optimizer with a learning rate of 0.0001, and regularization strategies such as EarlyStopping and ReduceLROnPlateau were applied to mitigate overfitting and adapt the learning rate dynamically. Table 2 presents the results in terms of test accuracy on FER2013 and the corresponding model size.

While ConvNeXt XLarge achieved the best classification accuracy (72.27%), its memory footprint (3.9 GB) exceeds the capacity constraints typically encountered in embedded or edge computing platforms.

4.1.2 Confidence Interval Estimation

Accuracy is an estimate of the system's performance, and its reliability depends on the number of evaluations—in our case, the number of facial expressions tested. To assess the statistical reliability of the reported recognition rates, we compute confidence intervals, following the method described in [63], which models successful predictions using a binomial distribution.

Let N be the number of test samples and P the recognition rate. The confidence interval at a confidence level $x\%$ (with z_x the corresponding critical value) is computed as:

[1] https://pal-robotics.com/robot/tiago/
[2] https://keras.io/api/applications/

Table 2 FER results using pretrained models fine-tuned on the FER2013 dataset: accuracy (%) and memory footprint (MB).

Pretrained model	Accuracy	FER system size
MobileNet	66.11	14.5
ResNet152V2	67.28	611.3
DenseNet201	67.84	221.0
InceptionV3	68.43	268.6
Xception	68.93	346.9
ConvNeXt Tiny	69.43	362
EfficientNetV2-B0	70.00	139.0
ConvNeXt Small	70.15	566
InceptionResNetV2	70.29	648.2
ConvNeXt Base	70.32	1120
VGG16	71.18	171.0
ResNet101	71.30	549.8
EfficientNetV2L	71.35	1448.7
VGG19	71.46	262.5
ConvNeXt Large	71.57	2733
ConvNeXt XLarge	**72.27**	**3900**

$$CI = \left[\frac{P + \frac{z_x^2}{2N} - z_x \sqrt{\frac{P(1-P)}{N} + \frac{z_x^2}{4N^2}}}{1 + \frac{z_x^2}{N}}, \frac{P + \frac{z_x^2}{2N} + z_x \sqrt{\frac{P(1-P)}{N} + \frac{z_x^2}{4N^2}}}{1 + \frac{z_x^2}{N}} \right] \quad (5)$$

where $z_{95\%} = 1.96$ and $z_{98\%} = 2.33$. This interval indicates that there is a probability of $x\%$ that the true recognition rate lies within $[P^-, P^+]$.
The FER2013 dataset contains 35,887 grayscale facial images, divided into training (80%), validation (10%), and test (10%) sets. Consequently, each model has been evaluated on 3,589 test samples. Using z_{98}, we compute the 98% confidence interval for all models and visualize the results in Figure 7.

We observe that several models—including VGG16, InceptionResNetV2, ConvNeXt Base, EfficientNetV2-B0, and VGG19—yield overlapping accuracy intervals. These models therefore exhibit statistically similar performance on FER2013. However, when considering both accuracy and model size, EfficientNetV2-B0 emerges as the most balanced option. It provides a compact architecture with a size of 139 megabytes and a total of 38.55 million parameters (including 7.10 million from the EfficientNetV2-B0 backbone and 31.45 million from the DNN), while maintaining performance that is statistically comparable to larger models.

4.2 Audiovisual emotion recognition

The objective of this study is to develop a lightweight and robust end-to-end Audiovisual Emotion Recognition (AVER) system. EfficientNetV2-B0 was selected for visual feature extraction due to its favorable balance between accuracy and model size. The subsequent challenge lies in designing an effective and compact fusion architecture

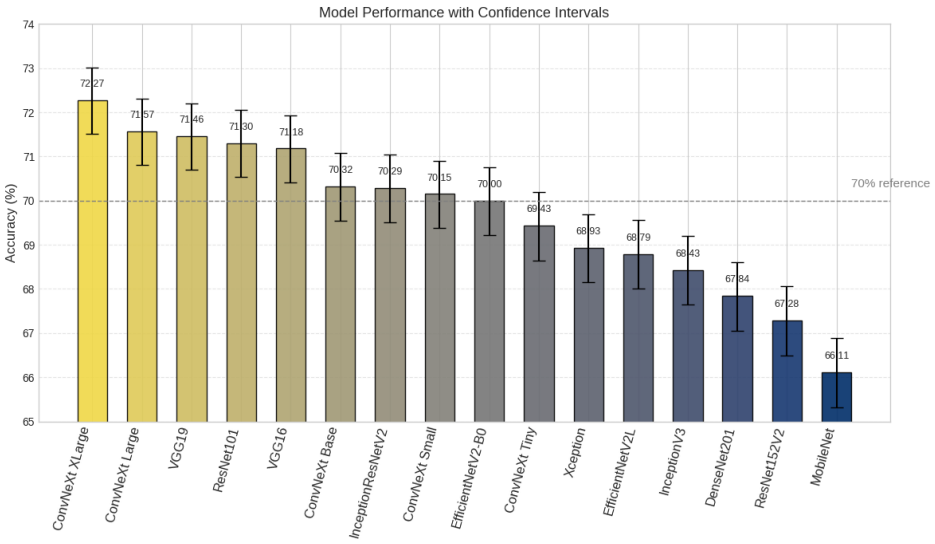

Fig. 7 Model performance assessment with 98% confidence intervals.

capable of integrating audio and visual modalities, while maintaining robustness in scenarios involving missing or unreliable input. To this end, we propose novel architectures that combine model-level fusion with late fusion mechanisms. These systems were systematically evaluated using the RAVDESS audiovisual dataset.

4.2.1 Dataset description

We use the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [39], which is a database specifically designed for the analysis and recognition of emotions from audio and video signals. It includes recordings from 24 actors, evenly divided between 12 men and 12 women, who pronounce the same phrases expressing the same set of emotions with the same number of repetitions, making the dataset balanced by design. The database covers eight distinct emotions: neutral, calm, happy, sad, angry, fearful, disgusted, and surprised (see Fig. 8). In this work, we use only the speech portion and not the sung recordings. These emotions are captured with two levels of intensity (normal and strong), and for some emotions, vocalized and non-vocalized versions are available. The data is provided in an audio-visual (mp4) format, thus offering an ideal framework for multimodal analysis.

4.2.2 Experimental protocol and results

Experimental protocol

The proposed framework is a strictly bimodal audiovisual emotion recognition system that integrates facial visual information and speech audio signals. Audiovisual recordings of 3.6 seconds are obtained by zero-padding the dataset segments when necessary.

Fig. 8 The RAVDESS dataset visual frames of some emotion categories [39]

The visual stream is represented as a sequence of 15 face-centered RGB frames uniformly sampled over time, forming a spatio-temporal input tensor of size $15 \times 224 \times 224 \times 3$. Faces are detected and cropped using the Multi-task Cascaded Convolutional Networks (MTCNN) algorithm [64], and the resulting images are resized to 224×224 pixels. For the selected images, the corresponding temporal instants were identified, and the associated 20 ms speech segments were extracted, each centered on its respective image. This temporal alignment between speech and facial data may reflect the presence of missing or incomplete information. This is particularly evident during silent pauses, where visual cues may still convey emotional content in the absence of corresponding audio signals. Furthermore, the process of temporal window normalization introduces zero-padding, which can also be interpreted as missing or non-informative data.

The feature extraction procedure was conducted as follows. The images were sequentially input into the pretrained model and passed through a dense layer of 128 neurons, yielding a 128-dimensional feature vector for each image. For the speech signal, 10 MFCCs, along with their first- and second-order derivatives, were extracted from each of the speech segments, with pitch and signal energy appended. Each segment vector was then concatenated with those of the two preceding and the following segments, resulting in audio vectors of 128 dimensions each.

For audiovisual fusion, separate Transformer-based cross-modal blocks are employed for Audio-to-Visual (AV) and Visual-to-Audio (VA) integration. Each block consists of an attention head of dimension 512 and a hidden layer of dimension 1024.

The overall model size is mainly driven by the pretrained EfficientNetV2-B0 backbone and the Transformer fusion modules, resulting in approximately 13.1 million parameters (7.1 million from EfficientNetV2-B0 and 6 million from the Transformer blocks). Variants incorporating additional self-attention mechanisms introduce approximately 1.8 million extra parameters, increasing the total model size to around 14.9 million parameters, while remaining suitable for deployment on resource-constrained devices.

Each actor in the RAVDESS dataset participated in 60 video sequences. The dataset was split into training, validation, and test sets using a strict subject-independent protocol, ensuring that actor identities do not overlap across splits. Specifically, 16 actors were used for training, 4 for validation, and 4 for testing, with results reported as the average over five cross-validation folds.

All models were trained for a maximum of 100 epochs using stochastic gradient descent (SGD) with a batch size of 64, a momentum of 0.9, and a weight decay of 10^{-3}. A ReduceLROnPlateau learning rate scheduler was employed with an initial learning rate of 0.04. Cross-entropy loss was used as the optimization objective, and early stopping was applied based on validation performance. All experiments were conducted using the PyTorch framework (v2.0.1) on an NVIDIA GeForce RTX 4090 GPU with 16 GB of memory.

Evaluation of the proposed architectures

Before comparing the proposed architectures with existing state-of-the-art methods, we first analyze the contribution of each modality within our audiovisual framework. In this setting, we train the baseline audiovisual architecture using both modalities. During evaluation, unimodal performance is obtained by fully masking one modality at the input—i.e., by zeroing out either the audio or the visual stream while keeping the other unchanged—without retraining the network. Using this protocol, the audio-only configuration achieves an accuracy of 58.08%, while the visual-only configuration reaches 72.83%. A quantitative comparison with previously reported baselines on RAVDESS further contextualizes these findings. In particular, Luna-Jiménez et al. [39, 48] proposed separate emotion recognition systems based on speech and facial cues. Their speech-based models achieved accuracies of 76.58% and 81.82%, while the face-based configurations reached 57.08% and 62.13%, respectively. They subsequently combined both modalities using a late fusion strategy to perform audiovisual emotion recognition. Similarly, the recent MultiMAE-DER model [35] reports 80.55% for audio-only and 74.13% for video-only performance. The audiovisual emotion recognition system is rather based on an early fusion. Compared to these works, our visual-only accuracy (72.83%) lies within the range of reported FER results, whereas our audio-only accuracy (58.08%) is lower than transformer-based SER approaches. This difference stem from several factors. For example, we opted to extract MFCC features for speech instead of using a pretrained model, with the aim of designing a lightweight architecture. In contrast, for the facial modality, feature extraction relies on an EfficientNet backbone, which may partly explain the observed performance gap. Furthermore, our framework is built upon an intermediate fusion strategy, which necessitates temporal synchronization between speech and facial streams. Consequently, silent segments were inevitably included during preprocessing, potentially affecting the overall performance. More importantly, our best multimodal configuration reaches 88.33%, yielding an improvement of +16.50 percentage points over the visual-only setting and +30.25 points over the audio-only setting. These gains are substantially larger than the multimodal improvements typically reported in prior studies, where the increase over the strongest unimodal modality generally ranges between 3 and 5 percentage points. This

observation further supports the complementary nature of audio and visual cues and highlights the effectiveness of the proposed fusion strategy.

Beyond this overall multimodal gain, we further analyze the contribution of each architectural refinement. The proposed architectures achieve accuracies of 84.37%, 87.29%, and 88.33% for the baseline, concatenation-based, and self-attention-based models, respectively, demonstrating progressive improvements across the designs, as shown in Fig. 9.

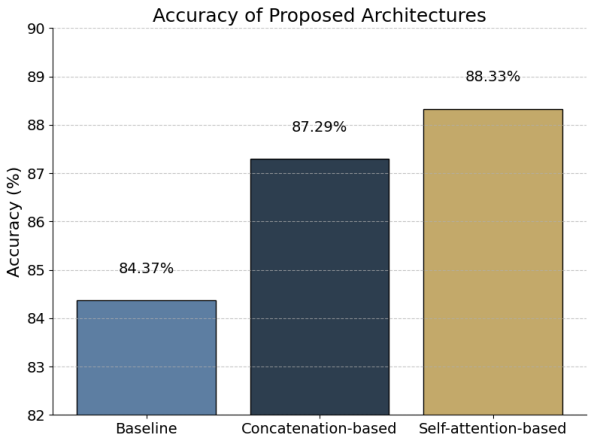

Fig. 9 Accuracy of the three proposed audiovisual architectures: baseline, concatenation-based, and self-attention-based architectures.

These results provide insight into the contribution of each architectural refinement within our framework.

- The baseline model already benefits from cross-modal Transformer blocks, which enable effective modeling of audio–visual correlations and yield competitive performance with a compact parameter budget.
- The concatenation-based variant further improves accuracy by explicitly preserving modality-specific representations alongside cross-modal features. This design choice enhances robustness in scenarios where inter-modal alignment is weak or partially missing, allowing each modality to contribute complementary emotional cues to the final prediction. Indeed, according to 4.2.2, our segments may contain some missing data, such as silent speech segments or missing faces due to the zero-padding process, making this design particularly beneficial.
- The self-attention-based architecture achieves the highest performance by additionally modeling intra-modal temporal dependencies before classification. By capturing long-range temporal dynamics independently within the audio and visual streams, the model is better equipped to exploit modality-specific patterns while maintaining effective cross-modal interaction. Indeed, our late fusion approach is based on the concatenation of three components: the outputs of the transformer blocks (2×128 parameters), the feature representations extracted from each modality (2×128 parameters), and the outputs of the self-attention blocks for each modality (2×128

parameters). Notably, the contributions of the three modalities are well balanced, ensuring that each modality contributes equitably to the final prediction.

Overall, the progressive improvement across architectures confirms that combining inter-modal attention, late fusion, and intra-modal self-attention yields more discriminative and robust emotion representations.

4.2.3 Model comparison

To further assess the effectiveness of our proposed audiovisual emotion recognition framework, we compare the proposed audiovisual architectures with a selection of recent state-of-the-art models evaluated on the same RAVDESS benchmark. These reference models represent a variety of multimodal fusion strategies, including handcrafted features, transformer-based architectures, and attention-guided fusion mechanisms. The comparison focuses on overall classification accuracy under subject-independent evaluation protocols, which provide a reliable indication of model generalization to unseen speakers.

Table 3 summarizes the accuracy scores reported in the literature alongside the results obtained by our three architectures: baseline, one based on simple feature concatenation and the other incorporating intra-modal self-attention mechanisms, together with their corresponding parameter counts when available.

Table 3 Performance comparison on the RAVDESS test set. Parameter counts are reported when available in the corresponding publications.

Model	Accuracy (%)	Params (M)
ERANN-0-4 [65]	74.8	24
CNN-14 & biLSTM-GuidedSTN[39]	80.08	81 (CNN-14 backbone)
Intermediate-Attention-Fusion [66]	81.58	Not reported
MultiMAE-DER [35]	83.61	Not reported
Ours: Baseline system	**84.37**	13.1
xlsr-Wav2Vec2.0 & bi-LSTM+Attention [48]	86.70	300
Ours: Concatenation-based approach	**87.29**	13.1
Ours: Self-Attention-based approach	**88.33**	14.9

These results highlight the effectiveness of our hybrid fusion strategy, which integrates inter-modal correlation modeling (via cross-attention mechanisms) with intra-modal late fusion structure learning (through concatenation and self-attention blocks). This combination enables the model to robustly capture emotion-relevant patterns even in cases of weak cross-modal alignment. Importantly, the proposed architectures achieve superior performance while relying on a significantly lower number of parameters compared to several competing approaches. For instance, our best-performing model reaches an accuracy of 88.33% with approximately 14.9 million parameters, whereas ERANN and CNN-14-based frameworks involve substantially larger models. In particular, the research presented in [48] reports a classification performance of 86%, which is close to ours; however, this model relies on a substantially larger number of parameters. This favorable trade-off between accuracy and model

compactness further supports the suitability of the proposed framework for real-time deployment on resource-constrained platforms.

Finally, the confidence intervals presented in Fig. 10 provide additional confirmation of this finding. It is evident that our two architectures, developed through combination with late fusion, achieve superior performance compared to existing systems.

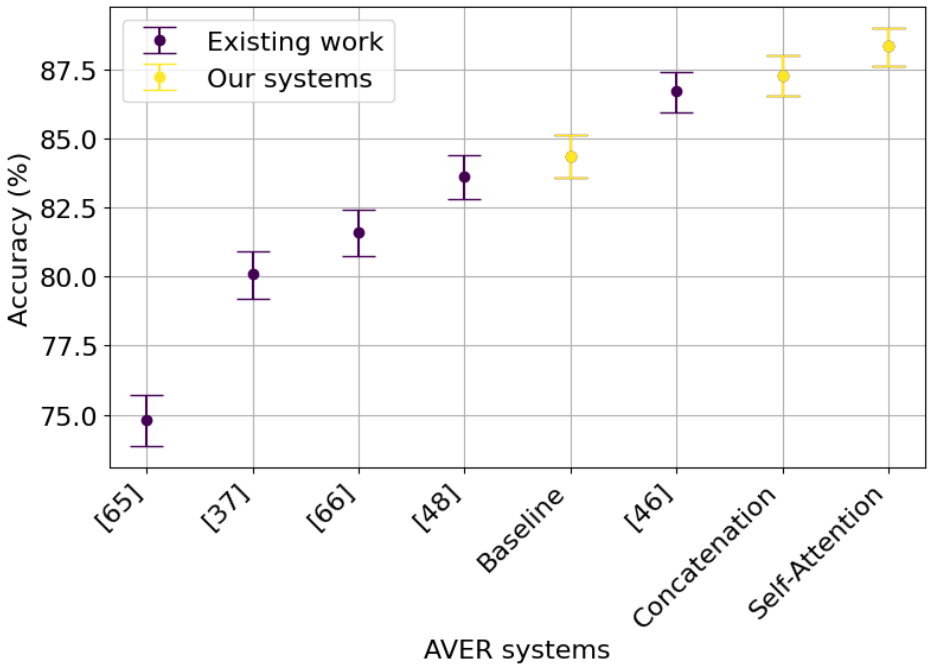

Fig. 10 Confidence intervals of our systems and existing systems evaluated on RAVDESS

5 Conclusions and perspectives

In this work, we introduced a lightweight, end-to-end audiovisual emotion recognition framework that effectively integrates speech cues and facial expressions using transformer-based deep learning architectures. To design an efficient end-to-end system, we first conducted a preliminary study to identify an optimal visual feature extractor. Through extensive benchmarking, EfficientNetV2-B0 was selected for its favorable trade-off between recognition performance and model complexity.

Building upon this foundation, we proposed three progressively refined audiovisual architectures that combine model-based fusion with late fusion strategies. These designs incorporate cross-modal attention mechanisms to model inter-modal correlations, as well as modality-specific pathways and intra-modal self-attention to enhance robustness in the presence of weak or missing modalities, while maintaining a compact model size.

Experimental results on the RAVDESS dataset demonstrate that the proposed architectures consistently outperform existing state-of-the-art methods while exhibiting a low memory footprint. These results highlight the effectiveness of the proposed fusion strategies and confirm their suitability for real-time deployment in resource-constrained environments such as embedded systems and mobile platforms.
Despite these promising results, RAVDESS remains a relatively small dataset based on acted emotional expressions. As a result, future work will focus on validating the proposed fusion strategies on larger-scale and more naturalistic multimodal datasets, such as Aff-Wild2, IEMOCAP, and more recent benchmarks including CG-MER and EMPATHIC. These datasets introduce greater variability, more spontaneous emotional behavior, and more challenging cross-modal synchronization conditions, enabling a more thorough assessment of generalization and real-world applicability.
Furthermore, although this study focused on audio and visual modalities, human emotions are inherently multimodal and can also be conveyed through additional cues such as body gestures and physiological signals. Extending the proposed framework to incorporate these modalities raises important technical challenges, including multimodal synchronization, increased data dimensionality, and robustness to missing or noisy signals. To address these challenges while preserving efficiency, future research will explore lightweight extensions of the proposed architecture based on modular modality-specific encoders, adaptive attention mechanisms that dynamically weight available modalities, and efficient temporal alignment strategies that avoid costly synchronization overhead. Such design choices aim to maintain a compact model footprint while enabling scalable multimodal integration suitable for real-time and embedded applications.

Declarations

5.1 Availability of Data and Materials

The datasets used in this study are publicly available. The FER2013 dataset can be accessed at https://www.kaggle.com/datasets/msambare/fer2013, and the RAVDESS dataset is available at https://zenodo.org/record/1188976

5.2 Competing Interests

The authors have no relevant financial or non-financial interests to disclose. There are no competing interests to declare that are relevant to the content of this article. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. Furthermore, the authors have no financial or proprietary interests in any material discussed in this article.

5.3 Funding

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

5.4 Authors' contributions

Both authors contributed to the research and preparation of the manuscript.

5.5 Acknowledgements

We would like to thank all participants for their contributions to this study. We also extend our gratitude to Mohamed Alaa Yahyaoui for his valuable efforts.

References

[1] Palmero, C., deVelasco, M., Hmani, M.A., Mtibaa, A., Letaifa, L.B., Buch-Cardona, P., Justo, R., Amorese, T., González-Fraile, E., Fernández-Ruanova, B., Tenorio-Laranga, J., Johansen, A.T., Silva, M.R., Martinussen, L.J., Korsnes, M.S., Cordasco, G., Esposito, A., El-Yacoubi, M.A., Petrovska-Delacrétaz, D., Torres, M.I., Escalera, S.: Exploring emotion expression recognition in older adults interacting with a virtual coach. IEEE Transactions on Affective Computing, 1–18 (2025)

[2] Rios, A., Reichel, U., Bhuvaneshwara, C., Filntisis, P., Maragos, P., Burkhardt, F., Eyben, F., Schuller, B., Nunnari, F., Ebling, S.: Multimodal recognition of valence, arousal and dominance via late-fusion of text, audio and facial expressions. In: Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2023) (2023)

[3] Justo, R., Letaifa, L.B., Palmero, C., Fraile, E.G., Johansen, A.T., Vazquez, A., Cordasco, G., Schlogl, S., Ruanova, B.F., Silva, M.R., Escalera, S., Velasco, M.D., Laranga, J.T., Esposito, A., Kornes, M., Torres, M.I.: Analysis of the interaction between elderly people and a simulated virtual coach. Journal of Ambient Intelligence and Humanized Computing **11**, 6125–6140 (2020)

[4] Wu, C.-H., Lin, J.-C., Wei, W.-L.: Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies. APSIPA Transactions on Signal and Information Processing **3**(1) (2014) https://doi.org/10.1017/ATSIP.2014.11

[5] Joze, H.R.V., Shaban, A., Iuzzolino, M.L., Koishida, K.: MMTM: Multimodal Transfer Module for CNN Fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)

[6] Ghimire, D., Lee, J., Li, Z.-N., Jeong, S.: Recognition of facial expressions based on salient geometric features and support vector machines. Multimedia Tools and Applications **76**, 7921–7946 (2017)

[7] Kotsia, I., Pitas, I.: Facial expression recognition in image sequences using geometric deformation features and support vector machines. IEEE transactions on image processing **16**(1), 172–187 (2006)

[8] Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: A comprehensive study. Image and vision Computing **27**(6), 803–816 (2009)

[9] Chen, J., Chen, Z., Chi, Z., Fu, H., *et al.*: Facial expression recognition based on facial components detection and hog features. In: International Workshops on Electrical and Computer Engineering Subfields, pp. 884–888 (2014)

[10] Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 1459–1462 (2010)

[11] Letaifa, L., Torres, M., Justo, R.: Adding dimensional features for emotion recognition on speech. In: Proceedings of the International Conference on Advanced Technologies for Signal and Image Processing (2021)

[12] Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech communication **53**(9-10), 1062–1087 (2011)

[13] Zhang, T.: Facial expression recognition based on deep learning: a survey. In: Advances in Intelligent Systems and Interactive Applications: Proceedings of the 2nd International Conference on Intelligent and Interactive Systems and Applications (IISA2017), pp. 345–352 (2018). Springer

[14] Li, S., Deng, W.: Deep facial expression recognition: A survey. IEEE transactions on affective computing **13**(3), 1195–1215 (2020)

[15] Issa, D., Demirci, M.F., Yazici, A.: Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control **59**, 101894 (2020)

[16] Letaifa, L.B., Torres, M.I.: Perceptual borderline for balancing multi-class spontaneous emotional data. IEEE Access **9**, 55939–55954 (2021)

[17] Bohi, A., Boudouri, Y.E., Sfeir, I.: A novel deep learning approach for facial emotion recognition: application to detecting emotional responses in elderly individuals with alzheimer's disease. Neural Computing and Applications **37**(6), 5235–5253 (2025)

[18] El Boudouri, Y., Bohi, A.: Emonext: an adapted convnext for facial emotion recognition. In: 2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP), pp. 1–6 (2023). IEEE

[19] Letaifa, L.B., Rouas, J.: Transformer model compression for end-to-end speech recognition on mobile devices. In: Proceedings of the 30th European Signal Processing Conference (EUSIPCO 2022), Belgrade, Serbia, pp. 439–443 (2022)

[20] Rouas, J., Brazier, C., Letaifa, L.B., Medina, R., Palacios, P., Atienza, D., Ansaloni, G.: Structured pruning for efficient systolic array accelerated cascade speech-to-text translation. In: Proceedings of the 26th Annual Interspeech Conference, Interspeech 2025 (2025)

[21] Oujabour, M., Letaifa, L.B., Dollinger, J., Rouas, J.: Adaptive compression of supervised and self-supervised models for green speech recognition. In: Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India (2025)

[22] Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., *et al.*: Challenges in representation learning: A report on three machine learning contests. In: Neural Information Processing. 20th International Conference, ICONIP, pp. 117–124 (2013). Springer

[23] Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2584–2593 (2017). IEEE

[24] Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017)

[25] Farzaneh, A.H., Qi, X.: Facial expression recognition in the wild via deep attentive center loss. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2402–2411 (2021)

[26] Pecoraro, R., Basile, V., Bono, V.: Local multi-head channel self-attention for facial expression recognition. Information **13**(9), 419 (2022)

[27] Han, B., Hu, M., Wang, X., Ren, F.: A triple-structure network model based upon mobilenet v1 and multi-loss function for facial expression recognition. Symmetry **14**(10), 2055 (2022)

[28] Fard, A.P., Mahoor, M.H.: Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. IEEE Access **10**, 26756–26768 (2022)

[29] Zhang, F., Chen, G., Wang, H., Zhang, C.: Cf-dan: Facial-expression recognition based on cross-fusion dual-attention network. Computational Visual Media, 1–16 (2024)

[30] Zheng, C., Mendieta, M., Chen, C.: Poster: A pyramid cross-fusion transformer network for facial expression recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3146–3155 (2023)

[31] Tian, Y., Zhu, J., Yao, H., Chen, D.: Facial expression recognition based on vision transformer with hybrid local attention. Applied Sciences **14**(15), 6471 (2024)

[32] Lu, X., Zhang, H., Zhang, Q., Han, X.: Multi-channel expression recognition network based on channel weighting. Applied Sciences **13**(3), 1968 (2023)

[33] Yahyaoui, M.A., Oujabour, M., Ben Letaifa, L., Bohi, A.: Multi-Face Emotion Detection for Effective Human-Robot Interaction. In: Proceedings of the 17th International Conference on Agents and Artificial Intelligence - Volume 1: ICAART, pp. 91–99 (2025). INSTICC

[34] Punuri, S.B., Kuanar, S.K., Kolhar, M., Mishra, T.K., Alameen, A., Mohapatra, H., Mishra, S.R.: Efficient net-xgboost: an implementation for facial emotion recognition using transfer learning. Mathematics **11**(3), 776 (2023)

[35] Xiang, P., Lin, C., Wu, K., Bai, O.: Multimae-der: Multimodal masked autoencoder for dynamic emotion recognition. In: 2024 14th International Conference on Pattern Recognition Systems (ICPRS), pp. 1–7 (2024). IEEE

[36] Tzirakis, P., Trigeorgis, G., Nicolaou, M.A., Schuller, B.W., Zafeiriou, S.: End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of selected topics in signal processing **11**(8), 1301–1309 (2017)

[37] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation **42**, 335–359 (2008)

[38] Kollias, D., Tzirakis, P., Nicolaou, M.A., Papaioannou, A., Zhao, G., Schuller, B., Kotsia, I., Zafeiriou, S.: Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. International Journal of Computer Vision **127**(6), 907–929 (2019)

[39] Luna-Jiménez, C., Griol, D., Callejas, Z., Kleinlein, R., Montero, J.M., Fernández-Martínez, F.: Multimodal emotion recognition on ravdess dataset using transfer learning. Sensors **21**(22), 7665 (2021)

[40] Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R.: Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing **5**(4), 377–390 (2014) https://doi.org/10.1109/TAFFC.2014.2336244

[41] Letaifa, L.B., Develasco, M., Justo, R., Torres, M.I.: First steps to develop a corpus of interactions between elderly and virtual agents in spanish with emotion. In: International Conference on Statistical Language and Speech Processing (2019)

[42] Justo, R., Letaifa, L.B., Olaso, J.M., López-Zorrilla, A., Develasco, M., Vázquez, A., Torres, M.I.: A spanish corpus for talking to the elderly. Conversational Dialogue Systems for the Next Decade, 183–192 (2021)

[43] Farhat, N., Bohi, A., Letaifa, L.B., Slama, R.: Cg-mer: a card game-based multimodal dataset for emotion recognition. In: Sixteenth International Conference on Machine Vision (ICMV 2023), vol. 13072, pp. 399–406 (2024). SPIE

[44] Ben Letaifa, L., Bohi, A., Slama, R.: The cg-mer dyadic multimodal dataset for spontaneous french conversations: Annotation, analysis and assessment benchmark. Journal on Multimodal User Interfaces (2025)

[45] Lian, Z., Li, Y., Tao, J., Huang, J.: Investigation of multimodal features, classifiers and fusion methods for emotion recognition. arXiv preprint arXiv:1809.06225 (2018)

[46] Zhang, S., Ding, Y., Wei, Z., Guan, C.: Continuous emotion recognition with audio-visual leader-follower attentive fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3567–3574 (2021)

[47] Lee, S., Han, D.K., Ko, H.: Multimodal emotion recognition fusion analysis adapting bert with heterogeneous feature unification. IEEE access **9**, 94557–94572 (2021)

[48] Luna-Jiménez, C., Kleinlein, R., Griol, D., Callejas, Z., Montero, J.M., Fernández-Martínez, F.: A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset. Applied Sciences **12**(1), 327 (2021)

[49] Goncalves, L., Busso, C.: Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features. IEEE Transactions on Affective Computing **13**(4), 2156–2170 (2022)

[50] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). Ieee

[51] Baltrušaitis, T., Ahuja, C., Morency, L.-P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 423–443 (2018)

[52] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008 (2017)

[53] Ben Letaifa, L., Rouas, J.-L.: Variable scale pruning for transformer model compression in end-to-end speech recognition. Algorithms **16**(9), 398 (2023) https://doi.org/10.3390/a16090398

[54] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for

mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)

[55] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)

[56] He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, pp. 630–645 (2016). Springer

[57] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[58] Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)

[59] Tan, M., Le, Q.: Efficientnetv2: Smaller models and faster training. In: International Conference on Machine Learning, pp. 10096–10106 (2021). PMLR

[60] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (2017)

[61] Karen, S.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556 (2014)

[62] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986 (2022)

[63] Zouari, L.: Vers le temps réel en transcription automatique de la parole grand vocabulaire. PhD thesis (2007)

[64] Gradilla, R.: Multi-task cascaded convolutional networks (mtcnn) for face detection and facial landmark alignment (2020). Accessed: 2020

[65] Verbitskiy, S., Berikov, V., Vyshegorodtsev, V.: Eranns: Efficient residual audio neural networks for audio pattern recognition. Pattern Recognition Letters **161**, 38–44 (2022)

[66] Chumachenko, K., Iosifidis, A., Gabbouj, M.: Self-attention fusion for audiovisual emotion recognition with incomplete data. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 2822–2828 (2022)