Grenze International Journal of Engineering and Technology, June Issue

An Exploration of Emotion Recognition using Deep Learning across Multiple Modalities: Spoken Language, Written Text, and Facial Expressions

Dhruva M.S[1], Sunitha R[2] and Chandrika J[3]
[1]Research Scholar, Dept Of CS&E, BNMIT, Banglore, VTU, Belagavi,India
Email: msdhruva@gmail.com
[2]Associate Professor, Dept Of AI&ML, BNMIT, Banglore, VTU, Belagvai, India
[3]Professor & head, Dept Of ISE, MCE, Hassan, VTU, Belagavi, India
Email: sunithar1389@gmail.com, jc@mcehassan.ac.in

***Abstract*— In human-computer interaction, emotion detection is essential because it allows computers to comprehend and react to humans' emotional states. In order to improve accuracy and robustness, this work offers a thorough examination into multimodal emotion identification using deep learning techniques. To get a more complex picture of human emotions, the research combines data from many modalities, including voice signals, physiological signals, and facial expressions. Convolutional Neural Networks (CNNs) are used in the suggested method to analyze facial expressions. A well-thought-out multimodal architecture fuses multiple modalities, enabling the model to recognize intricate relationships and patterns across many data sources. Extensive tests are carried out on multimodal emotion recognition benchmark datasets to assess the performance of the proposed technique. The results illustrate how much better the deep learning-based strategy is than more conventional techniques, and how well it can identify and distinguish between a variety of emotional states. Additionally, the model's resilience is evaluated in a range of settings, such as loud settings and culturally diverse contexts. This work offers a cutting-edge method for multimodal emotion identification, which advances the rapidly expanding area of affective computing. In addition to improving emotion identification systems' accuracy, the established framework shows off their potential for use in virtual reality, affective computing, and human-computer interaction. The results of this research open up new possibilities for comprehending and using human emotions in intelligent systems in the future.**

***Index Terms*— deep learning, fusion methods, MER (Multimodal Emotion recognition).**

I. INTRODUCTION

Accurately identifying multimodal emotions is a major advancement in affective computing, since combining data from several sources improves the breadth and precision of emotion analysis. Conventional emotion identification algorithms have limited capacity to represent the depth and complexity of human emotions since they often depend on unimodal input, such as speech patterns or facial expressions. Multimodal techniques, on the other hand, make use of a variety of data streams, including physiological signals, vocal intonations, and facial characteristics, to provide a more thorough understanding of emotional states. This all-encompassing method makes it possible to identify minute details and offers a more realistic portrayal of a person's emotional state in relation to their

Grenze ID: 01.GIJET.10.2.700

surroundings. The significant improvement achieved through multimodal emotion recognition holds immense implications for various domains. In human-computer interaction, systems can adapt dynamically to users' emotional states, personalizing responses and enhancing user experience. In healthcare, the ability to analyze physiological signals alongside behavioral cues contributes to more accurate and sensitive assessments of mental health and well-being. Furthermore, in virtual reality environments, multimodal emotion recognition can enhance immersion by enabling virtual characters to respond realistically to users' emotions, fostering a more engaging and authentic experience.

Overall, the recognition of multimodal emotions not only refines the precision of emotion analysis but also opens avenues for the development of more empathetic and context-aware technologies across diverse applications. This advancement represents a crucial step towards the seamless integration of emotional intelligence into intelligent systems, laying the groundwork for a future where technology can better understand and respond to the intricacies of human emotions.

Thus, researchers are increasingly using deep learning architectures like CNNs for image data, RNNs for sequence data, and attention mechanisms for feature weighting to harness multimodal data fusion. Deep learning and multimodal emotion identification allow more accurate and subtle emotion categorization and transdisciplinary applications including human-computer interaction, virtual reality, and mental health evaluation. This convergence underscores the transformative impact of deep learning in understanding and interpreting the complex tapestry of human emotions across diverse modalities.

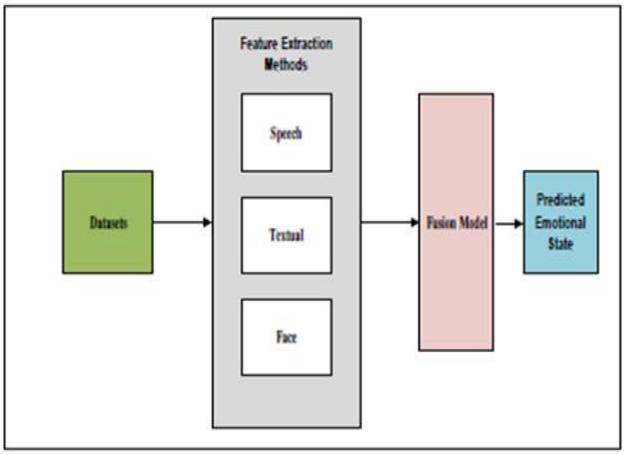

Figure 1. Standard framework for multimodal emotion recognition systems

In Figure 1, a standard framework deep learning-based multimodal emotion recognition system is described, forming the basis of our survey.

A. Multimodal-Emotion-Recognition Datasets

The goal is to undertake a thorough analysis of the materials that are already accessible.

1) IEMOCOP

The information captured by the Interactive Emotional Dyadic Motion Capture (IEMOCAP) comprises numerous modalities. Audio and video constitute the two primary modalities, capturing facial expressions and voice information, respectively.

Researchers [1] often use the IEMOCAP dataset to benchmark the performance of their multimodal emotion recognition models.

2) YOUTUBE DATASET

Morency et al. [2] introduced the "YouTube dataset" since YouTube is a vast platform with diverse content, and various datasets could be associated with it. However, if you're referring to datasets related to YouTube content, research, or analysis.

These datasets often include comment text, user information, timestamps, and other relevant details.

If there have been new developments or releases in YouTube-related datasets since my last update, recommended checking recent literature, research publications, or dedicated repositories for the most up-to-date information on YouTube datasets.

3) MOUD

It's possible[3] that this dataset was introduced or became more widely known after that date. Alternatively, it might be a specialized or domain-specific dataset that hasn't gained widespread recognition.

If have access to a specific source or documentation related to MOUD, should be able to find detailed information on the dataset's characteristics, modalities, purpose, and any associated tasks or benchmarks. This information may include details about the types of opinion utterances included, the modalities covered (e.g., text, audio, video), and any annotations or labels provided with the dataset.

If MOUD is a recent development, it may find relevant information by searching in academic databases, data repositories, or official project websites.

4) ICT-MMMO

The establishment of the Multi-Modal Movie Opinion (ICT-MMMO) database by Wollmer et al. [4]. For the latest information, it is recommended to refer to the official sources associated with the dataset, scholarly articles, or literature that was published after the specified date.

For detailed information on the ICT-MMMO database, recommending the academic papers, the official repository, or website associated with the dataset. Researchers often provide comprehensive documentation describing the dataset's construction, modalities included, annotation process, and potential use cases.

5) CMU-MOSI

Supporting multimodal context-based sentiment intensity research is the objective of the Multimodal Opinion-level Sentiment Intensity (CMU-MOSI). The multifaceted nature of opinion expression is captured by the CMU-MOSI dataset, which was made available for the Multimedia Grand Challenge at the 2018 ACM International Conference on Multimedia Retrieval (ICMR).

Researchers can use it to train and evaluate models for tasks such as predicting sentiment intensity in multimodal inputs.

II. Related Work

A 1-Dimensional Convolutional Neural Network (CNN) is used to extract unique speech characteristics like Wavegram and Wavegram-Logmel in a new technique for emotion identification utilizing audio and visual modalities that has been suggested in [5]. When it comes to forecasting impulsive and natural emotions, this strategy works better than both conventional and cutting-edge deep learning techniques.

The suggested method uses deep neural networks and correlation analysis to effectively use shared information for audio-visual emotion identification [6]. It is composed of a fusion network for emotion prediction and audio and visual networks to extract feature representations. Training involves a joint loss combining correlation and classification losses. Experimental results on multiple datasets demonstrate improved emotion recognition performance, indicating that common information enhances feature stability across modalities. However, limitations may arise from dataset-specific biases, generalization to real-world scenarios, and computational efficiency in real-time applications.

The study [7] reviews recent achievements in deep learning-based multimodal emotion recognition (DL-MER) integrating audio, visual, and text modalities. It discusses MER milestones, typical multimodal emotional datasets, deep learning model principles, current advances, state-of-the-art feature extraction and multimodal information fusion approaches, research obstacles, outstanding concerns, and future objectives. This description lacks particular deep learning models and methodologies and a full explanation of DL-MER research problems and outstanding questions. Recent advances in sensor and information technology allow robots to perceive and interpret human emotions via facial expressions, voice, behavior, and physiological signs [8]. Unimodality and multimodality sensors are key to affective computing emotion recognition research.

The suggested emotion identification system [9] uses deep learning from voice and video emotional Big Data. Convolutional neural networks (CNNs) receive voice mel-spectrograms and video frame representations. Dependence on training data quality and quantity may restrict system resilience and generalization.

The suggested approach [10] employs a transformer-based model with three branches—audio self-attention, video self-attention, and audio-video cross attention—to improve audiovisual emotion identification. Real-world applications may face audiovisual data quality issues, ambient noise, and generalization to varied emotional expressions and circumstances.

Modality-referenced 3D convolutional neural networks are used to simulate human emotion in this research [11]. It examines how various emotion recognition techniques work together. The study's [11] drawbacks may include

real-world implementation issues, reference modality selection biases, and the need for validation across varied groups and emotion manifestations. For better emotion identification, research has employed different modalities. This research [12] suggested audio sample augmentation and an emotion-oriented encoder-decoder to improve emotion identification.

Researchers [13] now use contextual information to predict emotions. The suggested technique detects visual relationships between the primary target and nearby items. Experimental findings reveal state-of-the-art CAER-S and competitive EMOTIC benchmark performance. Using datasets, feature extraction techniques, and MER algorithms, the review [45] covers current advances in deep learning-based multimodal emotion recognition (MER). The survey helps researchers choose appropriate methods and advances deep learning-based MER.

In [15], a novel bimodal speech emotion recognition system using linguistic and audio data is given. At the decision-making stage, it recommends combining sentiments and emotions from text and audio transcriptions. To ensure generalizability, the proposed bimodal system may need to be validated on a larger and more diverse dataset. The suggested multimodal emotion recognition system [16] used subject-wise 5-fold cross-validation to classify eight emotions with 80.08% accuracy on the RAVDESS dataset utilizing speech and face data. Among the study's limitations are the need for more research to address the mismatch issue with frame-based systems for tasks involving videos and the reliance on embedded information from pre-trained models for improved accuracy and robustness. A visual-audio emotion identification system that incorporates emotional components using machine learning techniques is proposed in the article [17]. For the visual portion, it develops a link between emotion category and dimension interval; for the audio component, deep convolutional neural networks are used to extract emotion-related data.

III. Multi Modal Fusion Methods

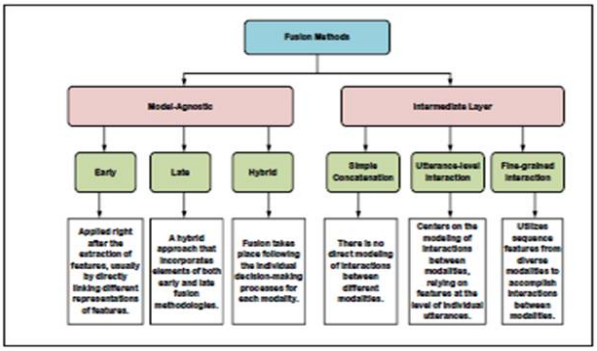

Fig 2. Multimodal Fusion methods

Nonverbal communication, with a special focus on motion, adds another degree of complication to the complicated work of emotion recognition. Unlike verbal cues, which are expressed through spoken words, nonverbal communication relies on subtle, often subconscious, cues embedded in facial expressions, body language, and gestures. Motion, as a key element of nonverbal communication, brings nuance to the understanding of emotions by incorporating dynamic aspects of human behavior. Decoding these nonverbal clues necessitates a deep exploration of psychological and physiological processes, delving into the realms of human cognition and emotion regulation.

Fusion techniques utilized in Multimodal Emotion Recognition (MER) based on deep learning can be classified into two main categories: intermediate layer fusion and model-agnostic fusion approaches. In model-agnostic fusion, separate unimodal models are trained for each modality (e.g., text, audio, and video), and their predictions are subsequently merged through a fusion procedure. This approach allows the flexibility of using different models for each modality and subsequently merging their outputs. On the other hand, intermediate layer fusion integrates information from different modalities at an intermediate layer within a shared neural network.

A. Model-Agnostic Fusion

In Multimodal Emotion Recognition (MER), model-agnostic fusion is a strategy in which different models are trained individually for each modality, such as text, audio, and video, and their predictions are then integrated via a fusion mechanism. This fusion method is agnostic to the underlying architecture of the individual models, allowing flexibility in choosing different models for each modality based on their specific characteristics.

1) Early Fusion

Poria et al. [18] proposed early fusion, a model-agnostic fusion technique in which input from many modalities is merged at the initial or "early" phases of a deep learning model. In the context of multimodal emotion recognition (MER). Huang et al. [19] improve the richness; this implies that characteristics from diverse modalities, such as text, audio, and video, are integrated before reaching the upper layers of the neural network. The primary idea is to create a joint representation that incorporates information from all modalities right from the input layer.
This joint representation is then used for subsequent layers of the network to learn complex patterns and relationships that exist across modalities.

2) Late Fusion

In the late fusion strategy, distinct models are trained for each modality individually, enabling them to capture modality-specific patterns and characteristics. After the individual modality models have processed their respective inputs by Poria et al. [20], the representations or predictions from each modality are combined or aggregated to make a final prediction for the overall task, such as emotion recognition.
Late fusion has various benefits. It enables for the use of diverse architectures for each modality, with each model optimized to capture the distinct properties of its input data. This approach is particularly useful when modalities have significantly different data structures or when specific processing is required for each modality. Late fusion can also be more computationally efficient, as the processing of each modality is distributed across separate branches.

3) Hybrid Fusion

A hybrid fusion technique, as proposed by Wollmer et al. [21], incorporates Bidirectional-Long-Short-Term-Memory (BLSTM). In the domain of MER, hybrid fusion is classified as model-agnostic fusion. It is a methodology that integrates components from early and late fusion strategies. And introduced a comparable hybrid fusion [22] process in their hybrid fusion architecture. The authors utilized a neural network with branches dedicated to independently processing individual modalities (late fusion aspect) and connections enabling the early fusion of specific features or representations from various modalities.

B. Intermediate Layer Fusion

Intermediate Layer Fusion, as an approach within model-agnostic fusion strategies in Multimodal Emotion Recognition (MER), involves integrating information from different modalities at intermediate layers of a neural network. Instead of fusing modalities at the input layer (early fusion) or at the output layer (late fusion), intermediate layer fusion allows for joint processing and feature extraction at mid-level stages of the network.

1) Utterance-Level Interaction Fusion

Utterance-Level Interaction Fusion is a model-agnostic fusion strategy in Multimodal Emotion Recognition (MER) that focuses on capturing interactions between modalities at the level of individual utterances or sentences. This approach aims to enhance the representation of multimodal data by considering the dynamic relationships and dependencies between different types of information within the context of a single expression or utterance.
Utterance-Level Interaction Fusion is particularly valuable in tasks where the relationships between modalities within short sequences of data (utterances or sentences) are critical such as in emotion recognition from conversational data or spoken language.
Designing effective interaction mechanisms and ensuring that they capture meaningful dependencies between modalities without introducing excessive complexity is a key challenge in this approach.
This fusion strategy provides flexibility by allowing the model to adaptively adjust the importance of different modalities based on the contextual dynamics of each utterance.
Utterance-Level Interaction Fusion enhances the expressive power of multimodal models by explicitly modeling interactions at the level of individual utterances. Researchers often explore variations of this approach to find the most effective mechanisms for capturing and utilizing inter-modal dependencies in the context of emotion recognition tasks.

2) Fain-Grained Interaction Fusion

It is essential to emphasize, nonetheless, that the effectiveness of these methodologies is contingent upon the accessibility of aligned annotations; the challenges linked to acquiring accurate alignments across modalities could potentially limit their applicability. Furthermore, in real-world applications, the computational cost of aligning features at fine-grained levels offers practical obstacles. Regardless of these issues, the development of fine-

grained interaction approaches demonstrates a dedication to improving the depth and accuracy of multimodal information fusion in emotion identification research. Continued study and breakthroughs in this field are expected to provide useful insights and enhancements to multimodal emotion identification systems.

To tackle the issue of cross-modal asynchrony, the Progressive Modality Reinforcement (PMR) technique was introduced by Lv et al. [23]. This approach utilizes message hubs and cross-modal Transformers to facilitate the exchange of information between modalities. In order to enhance sentiment prediction tasks, our iterative methodology progressively integrates modality-specific data and common information.

Furthermore, a textual architecture was proposed by Tzirakis et al. [24] that successfully incorporates various modal features through the utilization of Transformers and an attention-based fusion approach. By optimizing the interconnections between modalities, this architectural design significantly improves the efficacy of sentiment recognition.

IV. Conclusion

This comprehensive analysis examined the present state and historical progression of emotion identification technologies based on deep neural networks (DNNs). An analysis was conducted on widely used techniques for extracting features, comprehensive databases of multimodal emotions, and the influence of deep learning on this swiftly evolving domain. Considerable investigation was conducted into a range of fusion strategies, encompassing intermediate layer fusion, early fusion, late fusion, and hybrid fusion. Each of these techniques is crucial for the efficient integration of modal information into DNNs. Despite these advancements, significant obstacles continue to exist in the field of multimodal emotion recognition (MER), specifically concerning its integration into deep learning. Deep learning models' quality and generalization abilities are strongly reliant on datasets. An ideal dataset should be representative, diversified, and large enough to support high-quality labeling. The creation of such a dataset is required for efficient MER, your annotation of multimodal data is a time-consuming and costly operation that requires subjective judgment by specialists.

Due to temporal misalignment and feature heterogeneity, integrating data from diverse modalities offers issues. This intricacy frequently leads to inefficient use of supplemental information, affecting emotion recognition ability. By measuring information content, identifying useful modalities, and optimizing the fusion process, information theory and entropy give significant insights. MER necessitates an in-depth examination of personal data, posing ethical and privacy problems. Achieving a balance between model performance, user privacy, data utilization, and user rights is a critical issue in real-world applications.

To summarize, despite tremendous development in DNN-based MER, major challenges remain. Addressing these issues is critical not just for pushing the field's frontiers, but also for constructing robust, efficient, and ethical emotion identification systems. The future of MER contains significant progress and innovation potential, with an emphasis on overcoming obstacles and ushering in a new age in emotion recognition technology.

References

[1] Wöllmer, M.; Weninger, F.; Knaup, T.; Schuller, B.; Sun, C.; Sagae, K.; Morency, L.P. Youtube movie reviews: Sentiment analysis in an audio-visual context. IEEE Intell. Syst. 2013, 28, 46–53.

[2] Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.P. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv 2016, arXiv:1606.06259.

[3] Chou, H.C.; Lin, W.C.; Chang, L.C.; Li, C.C.; Ma, H.P.; Lee, C.C. NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus. In Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, TX, USA, 23–26 October 2017; IEEE: New York, NY, USA, 2017; pp. 292–298.

[4] Zadeh, A.B.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 2236–2246.

[5] T. Hussain, W. Wang, N. Bouaynaya, H. Fathallah-Shaykh and L. Mihaylova “Deep Learning for Audio Visual Emotion Recognition” IEEE 2022 DOI: 10.23919/FUSION49751.2022.9841342

[6] Fei Ma Wei Zhang Yang Li Shao-Lun Huang and Lin Zhang “Learning Better Representations for Audio-Visual Emotion Recognition with Common Information” Appl. Sci. 2020, 10(20), 7239; DOI: https://doi.org/10.3390/app10207239

[7] Shiqing Zhang, Yijiao Yang, Chen Chen, Xingnan Zhang, Qingming Leng and Xiaoming Zhao “Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects” ScienceDirect 2023 DOI https://doi.org/10.1016/j.eswa.2023.121692

[8] Cai Y, Li X and Li J. Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review. Sensors (Basel). 2023 Feb doi: 10.3390/s23052455.

[9] M. Shamim Hossain, Ghulam Muhammad, “Emotion recognition using deep learning approach from audio–visual emotional big data Information Fusion”, ScienceDirect 2018, https://doi.org/10.1016/j.inffus.2018.09.008.
[10] John, Vijay & Kawanishi, Yasutomo. “Audio and Video-based Emotion Recognition using Multimodal Transformers”. ICPR (2022). 10.1109/ICPR56361.2022.9956730.
[11] Noushin Hajarolasvadi, Hasan Demirel “Deep emotion recognition based on audio–visual correlation” IET 14 October 2020 DOI: https://doi.org/10.1049/iet-cvi.2020.0013
[12] Fu, Changzeng, Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro. 2020. "Multi-Modality Emotion Recognition Model with GAT-Based Multi-Head Inter-Modality Attention" Sensors 20, no. 17: 4894. https://doi.org/10.3390/s20174894 .
[13] Hoang, Manh-Hung & Kim, S.H. & Yang, Hyung-Jeong & Lee, Guee-Sang. “Context-Aware Emotion Recognition Based on Visual Relationship Detection”. IEEE Access. (2021). PP. 1-1. 10.1109/ACCESS.2021.3091169.
[14] Lian, Hailun, Cheng Lu, Sunan Li, Yan Zhao, Chuangao Tang, and Yuan Zong. 2023. "A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face" Entropy 25, no. 10: 1440. https://doi.org/10.3390/e25101440
[15] Oxana Verkholyak, Anastasia Dvoynikova1 and Alexey Karpov ”A Bimodal Approach for Speech Emotion Recognition using Audio and Text” Journal of Internet Services and Information Security (JISIS), volume: 11, number: 1 (February 2021), pp. 80-96 DOI:10.22667/JISIS.2021.02.28.080
[16] Luna-Jiménez.C, Griol.D, Callejas.Z, Kleinlein.R, Montero.J.M and Fernández-Martínez, F. “Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning”. Sensors 2021, 21, 7665. https://doi.org/10.3390/s21227665
[17] Jiajia Tian and Yingying She “A Visual–Audio-Based Emotion Recognition System Integrating Dimensional Analysis”. IEEE Transactions on Computational Social Systems. (2022). DOI: 10.1109/TCSS.2022.3200060
[18] Poria, S.; Cambria, E.; Hussain, A.; Huang, G.B. Towards an intelligent framework for multimodal affective data analysis. Neural Netw. 2015, 63, 104–116.
[19] Huang, J.; Li, Y.; Tao, J.; Lian, Z.; Niu, M.; Yang, M. Multimodal continuous emotion recognition with data augmentation using recurrent neural networks. In Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop, Seoul, Korea, 22 October 2018; pp. 57–64.
[20] Poria, S.; Cambria, E.; Howard, N.; Huang, G.B.; Hussain, A. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing 2016, 174, 50–59
[21] Wöllmer, M.; Weninger, F.; Knaup, T.; Schuller, B.; Sun, C.; Sagae, K.; Morency, L.P. Youtube movie reviews: Sentiment analysis in an audio-visual context. IEEE Intell. Syst. 2013, 28, 46–53.
[22] Nemati, S.; Rohani, R.; Basiri, M.E.; Abdar, M.; Yen, N.Y.; Makarenkov, V. A hybrid latent space data fusion method for multimodal emotion recognition. IEEE Access 2019, 7, 172948–172964.
[23] Lv, F.; Chen, X.; Huang, Y.; Duan, L.; Lin, G. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–25 June 2021; pp. 2554–2562.
[24] Tzirakis, P.; Chen, J.; Zafeiriou, S.; Schuller, B. End-to-end multimodal affect recognition in real-world environments. Inf. Fusion 2021, 68, 46–53.