AUXFORMER: ROBUST APPROACH TO AUDIOVISUAL EMOTION RECOGNITION

Lucas Goncalves and Carlos Busso

Multimodal Signal Processing (MSP) laboratory, Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA

goncalves@utdallas.edu, busso@utdallas.edu

ABSTRACT

A challenging task in audiovisual emotion recognition is to implement neural network architectures that can leverage and fuse multimodal information while temporally aligning modalities, handling missing modalities, and capturing information from all modalities without losing information during training. These requirements are important to achieve model robustness and to increase accuracy on the emotion recognition task. A recent approach to perform multimodal fusion is to use the transformer architecture to properly fuse and align the modalities. This study proposes the AuxFormer framework, which addresses in a principled way the aforementioned challenges. AuxFormer combines the transformer framework with auxiliary networks. It uses shared losses to infuse information from single-modality networks that are separately embedded. The extra layer of audiovisual information added to our main network retains information that would otherwise be lost during training. The results show that the AuxFormer architecture achieves macro and micro F1-Scores of 71.3% and 71.7%, respectively, on the CREMA-D corpus. For the MSP-IMPROV corpus, AuxFormer achieves a macro and micro F1-Scores of 70.4% and 76.5%, respectively. The results for both corpora are significantly better than strong baselines, indicating that our framework benefits from auxiliary networks. We also show that under non-ideal conditions (e.g., missing modalities) our architecture is able to sustain strong performance under audio-only and video-only scenarios, benefiting from a optimized training strategy.

Index Terms— Audiovisual emotion recognition, shared losses, multimodal fusion, transformers, auxiliary networks.

1. INTRODUCTION

The ability to accurately recognize human emotional states plays a key role in human communication [1]. When an interaction is taking place, being able to accurately determine a person's emotional state helps interlocutors to understand the message being conveyed better. Therefore, it is critical that modern *human computer interaction* (HCI) systems are designed to achieve similar performance by incorporating emotion recognition capabilities [2]. In real world situations, humans leverage acoustic and visual cues to determine the interlocutor's emotional state. For example, when a person speaks, expressive information is not only conveyed by speech features such as rhythm, pitch, intonation, and stress, but also by visual features like facial expressions. These features are considered synchronously during an interaction. It is important to implement a multimodal method that is able to process signals similarly to how human's perceive emotions in real world interactions.

Recent advances in deep learning approaches have achieved good performance in emotion recognition tasks when leveraging audiovisual features [3–5]. Audiovisual models still face challenging problems such as proper alignment between the modalities, robustness to missing modalities, and loss of information when fusing modalities. Studies have shown that audiovisual features are not synchronized, and there is a time-shift difference between the audio and visual streams. Differences between the modalities have been reported to be more than three video frames, which corresponds to one hundred milliseconds [6, 7]. Another important aspect of multimodal modeling is that emotion is dynamic and externalized overtime. Knowing an emotional state from previous time-steps/frames helps in predicting the current emotion. To ensure the model retains awareness to temporal emotional changes, it is essential to capture temporal information and certify that this information is not lost as we combine modalities. Likewise, during naturalistic interactions, a modality might be missing for certain periods of time (e.g., non audible speech, occluded faces). Previous studies have been presented to separately address these issues. Parthasarathy and Sundaram [3] used the transformer architecture to fuse modalities while introducing training strategies to handle missing modalities. Tsai et al. [8] proposed a multimodal transformer architecture for unaligned language sequences to fuse and align textual, visual, and acoustic features for emotion recognition. Addressing all these challenges is still an open problem.

Our objective is to implement an architecture that effectively handles temporal alignment of modalities, is robust to missing modalities, and does not lose important information from either modality throughout training. We chose a design that concurrently receives feature vectors from acoustic and visual modalities and fuses them together with cross-modal attention layers, while being able to temporally align relevant representations of each modality at the model level. Furthermore, we simultaneously train two auxiliary networks, which separately embed each modality into new vector spaces. The representations obtained from the auxiliary networks are used to provide an extra layer of information to our main network. These representations would otherwise be lost during training without the auxiliary networks. A shared losses mechanism is employed in our method to enable information to flow from the auxiliary networks into the main network. This idea was inspired by the study of Piergiovanni et al. [9] which explored a mechanism that learns multimodal representations by distilling information through losses from different modalities, and from the study of Szegedy et al. [10], which showed the benefits of auxiliary networks for very deep models. Our proposed method extends these concepts from real world situations to implement a training strategy that randomly eliminates different modalities during training to simulate situations where a person's head is occluded or a person's speech is inaudible during an interaction. We quantitatively evaluate and compare the performance of our proposed model against strong baselines, exploring a single-modality version of our architecture, and different training strategies of our architecture.

The results demonstrate that our architecture achieves significantly better results on both the CREMA-D corpus and the MSP-IMPROV corpus compared to the baselines under ideal conditions. Also, our architecture is able to sustain strong per-

This work was supported by NSF under Grant IIS-1718944

formance in non-ideal conditions when compared with audio-only and video-only architectures. These results show the benefits of our proposed AuxFormer architecture for audiovisual emotion recognition. The code for this work can be found at `https://github.com/ilucasgoncalves/AuxFormer`.

2. RELATED WORK

Model level fusion approaches have been more extensively explored with advancements of multimodal deep learning methods [11–13]. Studies such as the one done by Majumder et al. [14] have shown competitive performance in emotion recognition through a hierarchical learning mechanism, where unimodal, bimodal, and trimodal feature vector representations are generated from textual, visual, and acoustic features and fused at the model level. Recently, after the introduction of the transformer architecture [15], many studies have explored utilizing this architecture to implement multimodal transformer architectures for model level fusion [3–5]. These studies demonstrate strong performance by using a model level fusion approach, which allows the model to leverage information similarly to the way humans communicate during daily interactions.

Following the concepts presented by recent studies using attention for alignment [8, 16, 17], we took inspiration from these approaches to derive our architecture. Our novel approach aims to generate meaningful audiovisual latent representations that not only carry information from both modalities, but also carry temporal information which would have otherwise been lost. Our formulation removes the requirement of implementing extra steps to perform temporal alignment, in turn reducing biases present in manual alignment methods and delivering a more end-to-end framework.

3. RESOURCES

This study uses the CREMA-D [18] and MSP-IMPROV [19] corpora. CREMA-D is an audiovisual dataset, which contains videos of subjects saying sentences while displaying pre-defined emotional attributes aimed at demonstrating happiness, fear, disgust, anger, sadness and neutrality. This corpus was collected from an ethnically and racially diverse group consisting of 91 actors (48 male and 43 female). During the recording sessions, participants were given instructions to display certain emotional states while saying a target sentence. The obtained videos were annotated with emotional labels by 7 raters after watching the videos. In total, the CREMA-D corpus contains 7,442 clips with the following distribution: 1,222 disgust clips (11,429 ratings), 1,067 angry clips (10,054 ratings), 1,180 fear clips (11,153 ratings), 2,071 neutral clips (19450 ratings), 1,230 happy clips (11,730 ratings), and 672 sad clips (6347 ratings). This study, uses the six emotional states provided in the corpus.

The MSP-IMPROV corpus is an acted audiovisual dataset collected from recordings of dyadic interactions in improvised scenarios [19]. This corpus was created by using 20 target sentences that are meant to be said during improvised interactions, while targeting four different emotional states (sadness, happiness, anger, and neutrality) generating a total of 80 scenarios. The actors had to utter the target sentence during the improvisation. The corpus includes all the speaking turns during the interaction. It also includes natural interaction between breaks. The MSP-IMPROV includes 7,818 improvised interaction turns from 12 English speaking actors (6 males and 6 female). The emotional states from the MSP-IMPROV dataset were annotated using a crowd-sourced perceptual evaluation, using a protocol that tracks in real-time the quality of the workers as they complete their tasks [20]. The consensus labels are obtained using the plurality rule, obtaining 2,644 sentences for happiness, 792 sentences for anger, 3,477 sentences for neutral, and 885 sentences for sadness. A total of 85 speaking turns were labeled as other and 555 sentences did not reach an agreement. In our experiments, we only consider sentences labeled as happiness, anger, sadness, and neutral.

4. PROPOSED APPROACH

Fig. 1 shows an overview of the AuxFormer framework. The model is a deep learning architecture that uses the transformer as the main network to combine and align the audiovisual features. Along with our main network, we introduce single-modality auxiliary networks, which are simultaneously trained with the main network learning better representations through a shared loss mechanism.

4.1. Features and Pre-processing Steps

4.1.1. Visual Features

We start pre-processing our videos by retrieving every frame present in each video segment and estimating bounding boxes to retrieve faces from each frame. The bounding boxes are generated using a pre-trained face detector model developed by Liu et al. [21]. Following the face extraction step, the pixel intensities of the frames are normalized in the range [-1,1], resizing the images to 224x224x3. We then use the pre-trained VGG-face architecture [22] to obtain a feature vector of length 4,096 for each frame on the video segments. The frame-level feature vectors are row-wise concatenated into single matrices representing each video segment. After generating the feature vectors for each video segment, we make sure that all of the resulting vectors have the same dimensions by zero-padding the sequential feature vectors that did not share the same dimension.

4.1.2. Acoustic Features

We use the *OpenSmile* toolkit [23] to extract the feature set proposed for the paralinguistic challenge in Interspeech 2013 [24]. The input acoustic representation to our model is composed of 65 acoustic *low level descriptors* (LLDs) and its respective first order derivatives, generating a 130 dimensional feature vector. The LLDs are directly extracted from the raw audio for each video clip, which are pre-processed with a window length of 32 ms and a step size of 16 ms. All the features were then sequentially concatenated row-wise into a matrix which was subsequently z-normalized before using it as an input to our network. Lastly, we made sure that all the vectors had the same dimensions. We added zero padding on vectors that had unmatched dimensions to use the dot-product.

4.2. The AuxFormer Framework

The proposed AuxFormer framework consists of three networks: the main audiovisual fusion network $\mathcal{F}_{av}(\bullet)$, the auxiliary acoustic network $\mathcal{F}_{a}(\bullet)$, and the auxiliary visual network $\mathcal{F}_{v}(\bullet)$ (Fig. 1).

The main network of the AuxFormer is the audiovisual fusion network, which uses the transformer architecture to build pairs of bimodal representations that are formed by having Keys/Values of one modality interact with the Queries of a target modality. The framework receives the sequential audio ($x_a \in \mathbb{R}^{N_a \times 130}$) and visual ($x_v \in \mathbb{R}^{N_v \times 4,096}$) feature vectors. The first step is to project the visual features using a 1D convolution to match the shape of our acoustic feature vector, which becomes $x_v \in \mathbb{R}^{N_v \times 130}$. Following the visual feature transformation,we add positional encodings to the visual and acoustic vectors and feed the resulting vectors into the fusion attention layers. The fusion attention layers have the same structure as the transformer framework [15], except we cross-connect the query vectors across the network as in [3–5, 8]. The resulting representations obtained from the fusion attention layers are then input into a self-attention network which has the same basic structure of the fusion layers. However, the self-attention layers do not have the cross-layers query vectors. Furthermore, the self-attention layers are equipped with residual connections which have

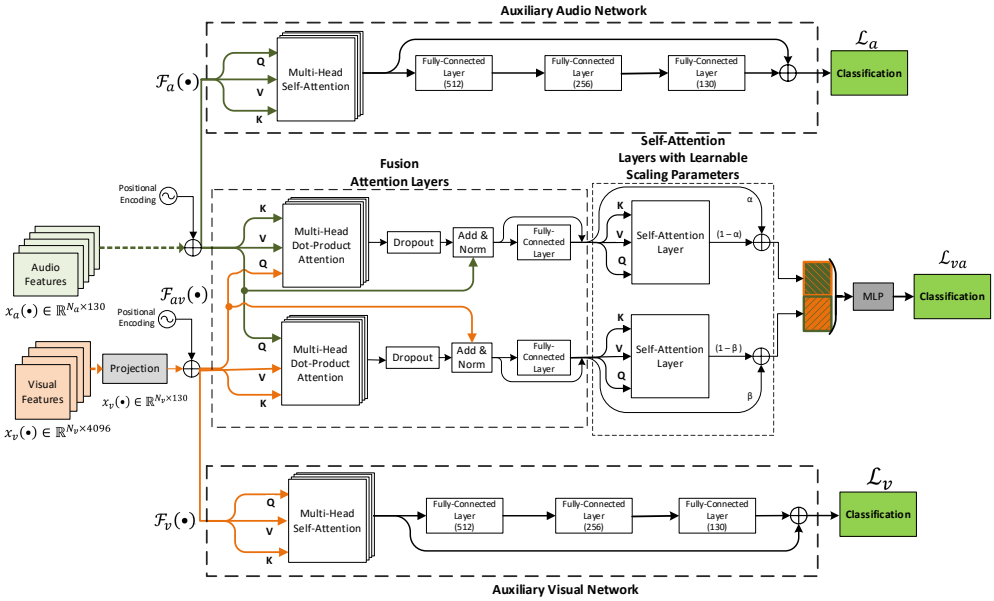


Fig. 1. The AuxFormer framework, which consists of the main audiovisual fusion network (middle) labelled $\mathcal{F}_{av}(\bullet)$, the auxiliary acoustic network (top) labelled $\mathcal{F}_a(\bullet)$, and the auxiliary visual network (bottom) labelled $\mathcal{F}_v(\bullet)$.

layer-independent learnable scaling weights α and β. A sequential classification token is then extracted from each resulting layer and concatenated before being fed into a *multilayer perceptron* (MLP) architecture for classification. The MLP layer we implemented after the concatenation of the sequential tokens obtained from the attention layers have two hidden layers of size 260, implemented with a *rectified linear unit* (ReLU) as the activation function. The size is 260 due to the shape of the concatenated tokens coming from each self-attention layer branch of dimension 130.

The auxiliary networks are identical with the exception that each network carries a specific modality (acoustic or visual). The auxiliary networks separately receive as input the sequential audio ($x_a \in \mathbb{R}^{N_a \times 130}$) and visual ($x_v \in \mathbb{R}^{N_v \times 130}$) feature vectors. The auxiliary network has a within modality multi-head self-attention layer, which generates embeddings to be processed through the fully-connected layers, as seen in Fig.1, for classification. The auxiliary networks also have residual connections across its fully-connected layers to ensure important information is not lost during training.

The training scheme of our proposed model makes use of shared losses to inject information into our main network from the auxiliary networks. Each network has its own loss $\mathcal{L}_m$, where $m \in \{a, v, av\}$ is the modality that the network is processing. We combine the training cross-entropy losses from each network during training to get the total loss used to train the model:

$$\mathcal{L}_{total} = \lambda_{va}\mathcal{L}_{va} + \lambda_a\mathcal{L}_a + \lambda_v\mathcal{L}_v, \quad (1)$$

where $\lambda_{va}, \lambda_a, \lambda_v$ are the weights for the specific losses of the audiovisual, audio, and visual networks.

4.3. Training Methods

Our goal with the proposed model is to achieve robust performance in non-ideal environments such as missing entire audio/video segments or having access to partial audio/video information only. We use two training procedures to evaluate our model's performance and robustness. First, we have a standard training, which consists of training our framework with the entire data available for training. Second, we have an optimized training procedure, which consists of randomly replacing the visual or the acoustic features layers with zeros with 20% probability each during training.

5. EXPERIMENTAL AND ARCHITECTURE EVALUATION

5.1. Experimental Settings

The multi-head attention layers and self-attention layers are all five layers deep, and each layer has 10 attention heads. We set a dropout rate of 0.25 to the output embeddings obtained from the attention layers, along with a 0.1 dropout rate for the residual connections. The model uses *adaptive moment estimation* (ADAM) as the optimizer with an initial learning rate set to 7.25E-04. During training, we have a learning decay set to 5 epochs. We set the gradient clipping threshold to 0.8, batch size to 32, and use ReLU as the activation function. $\lambda_{va}, \lambda_a, \lambda_v$ are set to have equal contribution (i.e. $\frac{1}{3}$). The model was trained for 20 epochs with an early stopping criterion based on the development loss. Everything was implemented in Pytorch and trained using a Nvidia QUADRO RTX 8000.

We compare our model with three baselines to verify our model's performance under ideal conditions against strong frameworks. Baseline 1 was derived from the AuxFormer framework. The only difference is that we removed the audio and the visual auxiliary networks from our framework. This baseline only uses the fusion attention layers, self-attention layers with learnable parameters, and MLP layer for emotion recognition. Baseline 2 was derived from the architecture presented in Tsai et al. [8] which proposed a multimodal transformer architecture for human language time-series data. We implemented their model making a few changes to their original architecture to adapt their method from three modalities (textual, visual, and acoustic) to two modalities (audiovisual). Baseline 3 was implemented using the transformer-based approach studied in Parthasarathy and Sundaram [3]. Furthermore, we analyze a single-modality version of our proposed model. All models were trained for 20 epochs. We also generate random splits in a speaker-independent manner amongst sets of 70% for training, 15% for development, and 15% for testing sets. We trained each model 20 times with different splits every time. We evaluated the models' performances using both

Table 1. Comparison of our model results with baseline audiovisual models. The table reports the average F-Score values across 20 experiments using different random seeds. (● indicates that one model is significantly better than baseline 1; ♦ indicates that one model is significantly better than baseline 2; ✖ indicates that one model is significantly better than baseline 3.)

	MSP-IMPROV		CREMA-D	
Architecture	Micro-F1	Macro-F1	Micro-F1	Macro-F1
AuxFormer	**0.765 ✖♦**	**0.704 ✖●♦**	**0.717 ✖●♦**	**0.713 ✖●♦**
Baseline 1	0.742	0.669	0.649	0.641
Baseline 2 [8]	0.737	0.666	0.614	0.604
Baseline 3 [3]	0.729	0.653	0.517	0.483

Table 2. Performance of our single-modality network architecture versions compared with the AuxFormer with optimized training model's performance on audio and visual only scenarios. The table reports the average F-Score values across 20 experiments using different random seeds. (♦ indicates that the model is significantly better than the Video Only architecture; ✖ indicates that the model is significantly better than the Audio Only architecture.)

	MSP-IMPROV		CREMA-D	
Architecture	Micro-F1	Macro-F1	Micro-F1	Macro-F1
AuxFormer (V)	0.721	0.598	**0.531 ♦**	**0.513 ♦**
Video Only	**0.736**	**0.652**	0.510	0.493
AuxFormer (A)	**0.624 ✖**	0.519	**0.587 ✖**	**0.578 ✖**
Audio Only	0.617	**0.539**	0.582	0.574

the macro-averaged F1-score and micro-averaged F1-score averaged across the 20 experiments performed on each model. Experimental results were recorded and compared using a one-tailed matched pair t-test over the 20 results with significance level at p-value <0.05 to assert statistical significance.

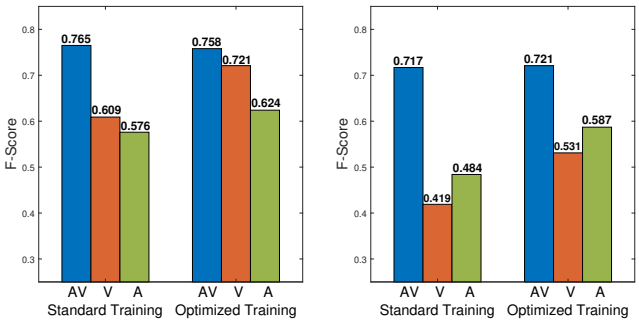
5.2. Evaluation with Baselines

In this section, we compare our proposed AuxFormer model's performance with the three baseline models. The models in this section were trained using standard training. Table 1 shows the F-scores of the models for the MSP-IMPROV and CREMA-D corpora. The AuxFormer architecture achieves the best F-Scores across both corpora, significantly outperforming all the baselines. We also notice, by comparing our model's performance with baseline 1, that the auxiliary networks play an important role in the performance of our model. This result supports our hypothesis that the auxiliary networks deliver an extra layer of important audiovisual information to our main architecture. Without the auxiliary networks we notice a significant drop in performance, as seen in Table 1.

5.3. Robustness Evaluation

In this section, we re-evaluate our model's performance under single-modality scenarios. We implement a unimodal version of our model and contrast it with the AuxFormer network under single-modality scenarios. We also evaluate the importance of performing optimized training procedures when training our framework.

We contrast the performance of our proposed model implemented with the optimized training method with the AuxFormer model implemented with the standard training method when we only have a single modality present during testing. Here, we clip the audio and video inputs to our networks by zeroing the acoustic and visual feature vectors, and individually record the performance with each modality on each corpus under standard and optimized training (Fig. 2). We notice that under the optimized training, the



(a) MSP-IMPROV Corpus (b) CREMA-D Corpus

Fig. 2. Graphs contrasting the performance of the AuxFormer model when using an standard training method versus the optimized training method to increase the model's robustness in non-ideal scenarios.

AuxFormer architecture outperforms the standard training network by 11.2% for video-only scenarios and by 4.8% for audio-only scenarios on the MSP-IMPROV corpus. It outperforms the standard training network by 11.2% for the video-only scenarios and by 10.3% for the audio-only scenarios on the CREMA-D corpus, while keeping strong performance in audiovisual scenarios.

To verify how our model performs under non-ideal conditions, we set up a unimodal baseline version of our model. In the unimodal version of our architecture, we replace the fusion attention layers with self-attention layers without learnable parameters and remove the auxiliary branch of the second modality from the framework. The unimodal network is then trained and tested with one modality at a time. Table 2 shows the performance of the unimodal version of our model, where the audio network performance is represented by the "Audio Only" entry and the visual network performance is represented by the "Video Only" entry. Table 2, shows that the AuxFormer network trained with the optimized training mechanism achieves better or consistently close performance to our single-modality baseline in audio-only and video-only scenarios for both corpora used in this study. The optimized training mechanism combined with auxiliary networks adds robustness to our model by enabling our model to learn from scenarios that simulate situations where a modality might be missing during test time, ensuring that our model does not lose important information from either of the modalities available during training.

6. CONCLUSIONS

We presented the AuxFormer framework, a new robust approach to audiovisual emotion recognition for non-ideal scenarios. The AuxFormer approach is able to perform audiovisual emotion recognition that effectively handles temporal alignment of modalities, is robust to missing modalities, and does not lose important information from either modality throughout training. We achieve this goal by using the transformer architecture and auxiliary networks with a shared losses mechanism that enables information to flow from the auxiliary networks to the main audiovisual network. By using a training strategy which simulates scenarios where certain modalities are missing during training, we achieve further performance robustness. Future work for this study is to explore self-supervised methods for pre-training our framework aiming to strengthen the model's performance and robustness against all possible scenarios. A promising future direction is to incorporate this framework to solve tasks outside of the emotion recognition tasks, such as audiovisual automatic speech recognition [25, 26].

7. REFERENCES

[1] K.R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, April 2003.

[2] N. Fragopanagos and J.G. Taylor, "Emotion recognition in human-computer interaction," *Neural Network*, vol. 18, no. 4, pp. 389–405, May 2005.

[3] S. Parthasarathy and S. Sundaram, "Training strategies to handle missing modalities for audio-visual expression recognition," in *International Conference on Multimodal Interaction (ICMI 2020)*, Utrecht, The Netherlands, October 2020, pp. 400–404.

[4] W. Rahman, M.K. Hasan, S. Lee, A.B. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Association for Computational Linguistics (ACL 2020)*, Online, July 2020, pp. 2359–2369.

[5] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 2803–2807.

[6] T.J. Hazen, "Visual model structures and synchrony constraints for audio-visual speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 1082–1089, May 2006.

[7] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, September 2003.

[8] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Association for Computational Linguistics (ACL 2019)*, Florence, Italy, July 2019, vol. 1, pp. 6558–6569.

[9] A. Piergiovanni, A. Angelova, and M. S. Ryoo, "Evolving losses for unsupervised video representation learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*, Seattle, WA, USA, June 2020, pp. 130–139.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, Boston, MA, USA, June 2015, pp. 1–9.

[11] Z. Zeng, J. Tu, B.M. Pianfetti, and T.S. Huang, "Audio-visual affective expression recognition through multistream fused HMM," *IEEE Transactions on Multimedia*, vol. 10, no. 4, pp. 570–577, June 2008.

[12] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Emotion recognition in the wild with feature fusion and multiple kernel learning," in *International Conference on Multimodal Interaction (ICMI 2014)*, Istanbul, Turkey, November 2014, pp. 508–513.

[13] G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M.I. Jordan, "Learning the kernel matrix with semidefinite programming," *Journal of Machine Learning Research*, vol. 5, pp. 27–72, January 2004.

[14] N. Majumder, D. Hazarika, A.F. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowledge-Based Systems*, vol. 161, no. 1, pp. 124–133, December 2018.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *In Advances in Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, December 2017, pp. 5998–6008.

[16] F. Tao and C. Busso, "Aligning audiovisual features for audiovisual speech recognition," in *IEEE International Conference on Multimedia and Expo (ICME 2018)*, San Diego, CA, USA, July 2018, pp. 1–6.

[17] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Association for Computational Linguistics (ACL 2018)*, Melbourne, Australia, July 2018, pp. 2225–2235.

[18] H. Cao, D.G. Cooper, M.K. Keutmann, R.C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, October-December 2014.

[19] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, January-March 2017.

[20] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A.C. Berg, "SSD: Single shot multibox detector," in *European Conference on Computer Vision (ECCV 2016)*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9905 of *Lecture Notes in Computer Science*, pp. 21–37. Springer Berlin Heidelberg, Amsterdam, the Netherlands, October 2016.

[22] O.M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference (BMVC 2015)*, Swansea,UK, September 2015, pp. 1–12.

[23] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[24] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.

[25] F. Tao and C. Busso, "End-to-end audiovisual speech recognition system with multi-task learning," *IEEE Transactions on Multimedia*, vol. 23, pp. 1–11, January 2021.

[26] F. Tao and C. Busso, "Gating neural network for large vocabulary audiovisual speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1286–1298, July 2018.