LEARNING CROSS-MODAL AUDIOVISUAL REPRESENTATIONS WITH LADDER NETWORKS FOR 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

Representation learning is a challenging, but essential task in audiovisual learning. A key challenge is to generate strong cross-modal representations while still capturing discriminative information contained in unimodal features. Properly capturing this information is important to increase accuracy and robustness in audiovisual tasks. Focusing on emotion recognition, this study proposes novel cross-modal ladder networks to capture modality-specific information while building strong cross-modal representations. Our method utilizes representations from a backbone network to implement unsupervised auxiliary tasks to reconstruct intermediate layer representations across the acoustic and visual networks. The skip connections between the cross-modal encoder and decoder provide powerful modality-specific and multimodal representations for emotion recognition. Our model on the CREMA-D corpus achieves high performance with precision, recall, and F1 scores over 80% on a six-class problem.

Index Terms— representation learning, audiovisual emotion recognition, ladder networks.

1. INTRODUCTION

Emotion perception plays an important role in human communication [1]. During naturalistic interactions, we rely on the externalization of acoustic and visual cues to perceive and convey emotional states. Hence, modern *human computer Interaction* (HCI) systems should be able to effectively utilize the same cues used by humans to predict the emotional state of users. Recent advancements in audiovisual emotion recognition systems have shown improvement over unimodal emotion recognition models. Developing methods to exploit acoustic and visual features ensuring that the relationship between both modalities is properly represented is fundamental to building robust multimodal systems.

Audiovisual emotion recognition has attracted increasing attention in recent years [2–5]. Although these models have shown strong performances, multimodal models still face major challenges [6]. In audiovisual settings, models have to process data points coming from heterogeneous sources (e.g., microphone and camera). Obtaining good representations from multimodal data points is essential for the performance of audiovisual systems. We argue that strong representations should capture the intrinsic and dynamic connection between facial expressions with acoustic features [7]. Existing representation learning methods can be grouped into two groups: forward and backward methods [8]. In forward methods, studies explore the use of cross-modal architectures, which focus on building interactions between different modalities [4, 9–11]. Since these methods are mainly focused on generating cross-modal representations, they often lack the capability of capturing unimodal information from the input data. In backward methods, studies utilize extra loss functions as a prior constraint to obtain complementary information between the modalities [8, 12]. These methods often utilize extra unimodal labels or extra unimodal encoders to capture modality-specific representations. However, these approaches are often costly, since they require more labels, and their feature representations are often not as strong in capturing cross-modal information.

This paper proposes a novel framework that is capable of capturing modality-specific information while building strong cross-modal representations without requiring extra human annotations. The proposed method incorporates both forward and backward techniques for multimodal representation learning. The framework utilizes representations from a backbone network to implement unsupervised auxiliary tasks with multimodal ladder networks, which utilize skip connections between the encoder of one modality and the decoder of the other modality, learning powerful modality-specific and cross-modal representations. The ladder network structure is used to combine our primary supervised task with unsupervised auxiliary tasks. The auxiliary task explored in this model is the reconstruction of hidden representations from one modality (e.g., acoustic) to another (e.g., visual) of a denoising autoencoder.

We compare our model with three state-of-the-art audiovisual baselines. The results demonstrate that our architecture achieves significantly better results on a six-class problem on the CREMA-D corpus [13]. Our multimodal ladder network framework achieves precision, recall, and F1 scores over 80%. An ablation analysis demonstrates the importance of the cross-modal reconstruction loss imposed by the proposed multimodal ladder network. While ladder network has been successfully used for *speech emotion recognition* (SER) [14–18], this is the first time, to the best of our knowledge, that ladder networks are used to create audiovisual representations.

2. RELATED WORK

Multimodal representation learning approaches have been explored more extensively lately with advancements in computational resources and multimodal deep learning methods [4, 5, 10, 12, 19, 20]. Previous studies have achieved major improvements on multi-modality representation learning. In multi-modality fusion, studies such as the one done by Majumder et al. [21] have shown competitive performance in emotion recognition through a hierarchical learning mechanism. Zadeh et al. [20], proposed using a tensor fusion network to aggregate representations from unimodal, bimodal and trimodal interactions. Yu et al. [8] proposed a late fusion framework, which uses self-supervised strategies to jointly learn unimodal and multimodal tasks to perform a multimodal emotion sentiment analysis regression task.

More recently, with the introduction of the transformer architecture [22], several studies have explored using this architecture for learning model-level cross-modal representations [3, 11]. Tsai et al. [10], proposed the use of cross-modal transformers to learn

This work was supported by NSF under Grant IIS-1718944

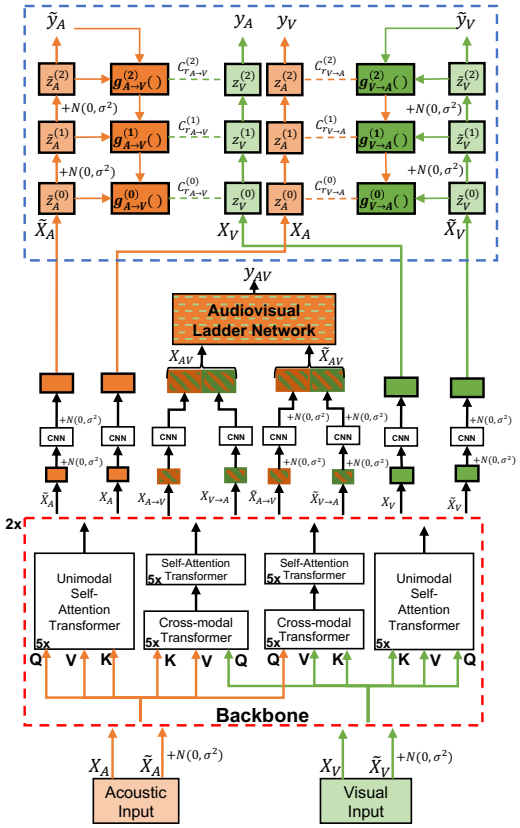

Fig. 1. Overview of the proposed multimodal ladder network. The red dashed rectangle shows the transformer-based backbone, which is inspired by the AuxFormer structure [4, 5]. Noisy and clean outputs from these layers are passed to CNN layers, which relay outputs to the ladder networks. On the top of our model, we implement the audiovisual ladder networks and the unimodal cross-layers ladder networks (shown inside the dashed blue rectangle).

multimodal representations from attention computations performed across different modalities within the transformer layers. Goncalves and Busso [4, 5], proposed the AuxFormer framework which uses a cross-modal transformer model with unimodal networks to build strong audiovisual fusion. The network also has unimodal networks to capture discriminative information from each modality.

3. PROPOSED APPROACH

This study proposes an audiovisual representation learning framework that combines the use of a backbone network containing transformer layers and the use of a novel ladder network framework that reconstructs cross-modal structures of an encoder-decoder network. Figure 1 provides an overview of the proposed model. We use the backbone network to implement unsupervised auxiliary tasks with our proposed multimodal ladder networks, which utilize skip connections between the encoder of one modality (e.g., acoustic) and the decoder of the other modality (e.g., visual). This method learns audiovisual representations by combining our primary supervised task with the unsupervised auxiliary tasks of reconstructing cross-modal representations of a denoising autoencoder.

3.1. Transformer-Based Backbone

The red dashed region in Figure 1 shows the backbone structure, which is inspired by the framework proposed by Goncalves and Busso [4, 5]. The architecture has cross-modal networks and unimodal networks. The cross-modal networks are composed of two separate cross-modal transformer layers that computes attention scores for the audiovisual modalities with sequence of length N. The input feature vectors are used to generate the Q, V, and K matrices. To inject information from one modality to the other, we enter the Q matrix from one modality (e.g., acoustic) at the cross-modal layer of the other modality (e.g., visual), and vice-versa as shown in Figure 1 and Equation 1.

$$\mathbf{z}_{aud \rightarrow vid} = softmax(\frac{Q_{vid} K_{aud}^{\top}}{\sqrt{N_{aud}}}) V_{aud} \quad (1)$$

The representations obtained from each cross-modal layer are subsequently processed by self-attention transformer layers, as shown in Figure 1. The combination of cross-modal transformer and self-attention transformer is five layers deep, generating the audiovisual representation vectors $X_{A \rightarrow V}$ and $X_{V \rightarrow A}$. In addition to the cross-modal layers, the backbone network contains two identical unimodal networks, one for the acoustic features and one for the visual features. These unimodal networks are important, since they ensure that modality-specific information is captured by the models, making it available for our multimodal ladder networks. The unimodal networks are implemented with five self-attention transformer layers, creating the visual representation vector X_V, and acoustic representation vector X_A.

The ladder network requires a clean and noisy version for each of the feature representations (i.e., X_A, X_V, $X_{A \rightarrow V}$ and $X_{V \rightarrow A}$). We create two versions for each input: a clean version, and a noisy version where we add Gaussian noise with variance σ^2. The block shown inside the dashed red region in Figure 1 is implemented twice, one for the clean inputs and one for the noisy inputs. The noisy feature representations are denoted by $\tilde{X}_A$, $\tilde{X}_V$, $\tilde{X}_{A \rightarrow V}$ and $\tilde{X}_{V \rightarrow A}$.

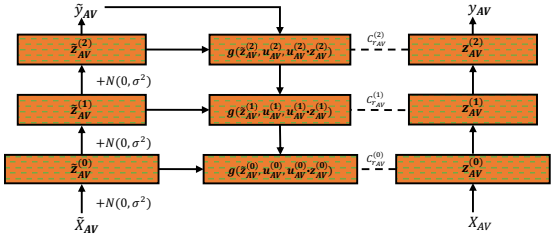

Fig. 2. Overview of the basic structure of the ladder networks. The audiovisual ladder network block displayed in the orange box in Figure 1 is implemented with this model using X_{AV} and $\tilde{X}_{AV}$ as input.

3.2. CNN Layers

The eight outputs from the transformer-based backbone are all processed through 1D temporal *convolutional neural networks* (CNNs) to ensure that the encoded sequences are aware of their neighboring components. This block is composed of three CNN layers with three max-pooling layers in between. As shown in Figure 1, we add Gaussian noise to the noisy feature representations $\tilde{X}_A$, $\tilde{X}_V$, $\tilde{X}_{A \rightarrow V}$ and $\tilde{X}_{V \rightarrow A}$. The CNN layers generate a sentence-level representation by flattening the CNNs outputs.

3.3. Ladder Networks

We utilize ladder networks to combine our primary supervised task with unsupervised auxiliary tasks. The auxiliary task consists of reconstructing hidden representations of a denoising autoencoder. Before explaining our proposed architecture, we describe the general form of the ladder network, which is shown in Figure 2. The ladder network contains lateral connections between the encoder and the

decoder layers, which allows the decoder to directly learn representations from the encoder layers.

Encoder: The ladder networks' encoder is composed of a fully-connected *multilayer perceptron* (MLP) network. As done previously, Gaussian noise with variance σ^2 is added to each layer of the noisy encoder (Fig. 2). The representation from the final layer $\tilde{z}^{(L)}$ of the encoder is used as a target for the supervised task. The decoder's goal is to reconstruct the representation at every layer, using a clean copy of the encoder **z** as the target.

Decoder: The decoder's goal is to denoise the encoded noisy layers. The denoising function, $g()$, shown in the mid layers in Figure 2, combines information from the decoder coming from the top layers and the side connections from the relative encoder layer. The structure of the ladder networks enables the top layers to learn more discriminative representations for our task, while the lower layers are focused on input vector reconstruction. The denoising function used in our ladder networks utilizes the function proposed by Pezeshki et al. [23], which is composed of an MLP with inputs $[u, \tilde{z}, u \odot z]$, where u is the batch normalized projection of the previous top layer. The overall loss function of the ladder networks is defined as:

$$C = C_s + \lambda_l \sum_l C_r^{(l)} \tag{2}$$

where C_s is the supervised loss, $C_r^{(l)}$ is the reconstruction loss at layer l, and λ_l is the loss weight hyperparameter.

3.4. Multimodal Ladder Network

As seen in the top region of Figure 1, we have two different settings for our multimodal ladder networks: the unimodal cross-layer ladder networks (shown inside the dashed blue rectangle in Fig. 1), and the audiovisual ladder network (shown in a orange box in Fig. 1). Although, we have two different settings, the overall structure of each setting is the same as the model described in Figure 2.

The unimodal cross-layer ladder networks combine the encoder from one modality with the decoder from the other modality. As shown in Figure 1, the inputs of the models are (X_A, $\tilde{X}_A$) for the acoustic modality, and (X_V, $\tilde{X}_V$) for the visual modality. These inputs only have information from the corresponding modality in the backbone networks. This strategy corresponds to a backward approach during fusion. In our proposed model, we generate noisy encoded representations from one modality $\tilde{z}_{M1}^{(L)}$. Then, the decoder has to reconstruct the representation at every layer, using clean encoded representations from a second modality z_{M2} as the target. In this case, the denoising function used in our ladder network inputs is defined as $[u_{M2}, \tilde{z}_{M1}, u_{M2} \odot z_{M2}]$. This procedure is done to explore the complementary relationship contained in audiovisual features, and help build strong cross-modal representations from modality-specific layers. It is done twice with acoustic features as noisy features and visual features as clean targets and vice-versa.

The audiovisual ladder network takes as input $\tilde{X}_{AV}$ and $\tilde{X}_{AV}$, which are processed by the cross-modal transformer. The structure of the ladder network is the same as the general structure shown in Figure 2. This procedure is important in our framework, because it is focused only on ensuring that cross-modal representations are adequately captured by the system.

3.5. Loss Function

For our entire model, we expand the loss function from Equation 2 to include all the objectives used in our model. Our overall loss is:

$$\mathcal{L}_{sup} = \frac{1}{3}(C_{s_{AV}} + C_{s_A} + C_{s_V}) \tag{3}$$

$$\mathcal{L}_{uns} = \lambda_l \left(\sum_l C_{r_{AV}}^{(l)} + \frac{1}{2} \left(\sum_l C_{r_{A\to V}}^{(l)} + C_{r_{V\to A}}^{(l)} \right) \right) \tag{4}$$

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \mathcal{L}_{uns} \tag{5}$$

where $\mathcal{L}_{sup}$ is the overall supervised loss, which combines the three supervised losses as the average of the supervised losses obtained from the three ladder networks composing our model. $\mathcal{L}_{uns}$ is the overall unsupervised loss, which consists of the combination of the unsupervised audiovisual loss $C_{r_{AV}}^{(l)}$ and the average of the two unsupervised losses obtained from the unimodal cross-layer predictions $C_{r_{A\to V}}^{(l)}$ and $C_{r_{V\to A}}^{(l)}$. The total loss $\mathcal{L}_{total}$ is the combination of $\mathcal{L}_{sup}$ and $\mathcal{L}_{uns}$.

4. RESOURCES AND FEATURES

4.1. Resources

This study uses the CREMA-D [13] corpus, which is an audiovisual dataset. It contains videos of subjects saying sentences while expressing pre-defined emotions (happiness, fear, disgust, anger, sadness, and neutral state). This corpus was collected from an ethnically and racially diverse group consisting of 91 actors (48 male and 43 female). The videos were annotated with emotional labels by seven raters using a crowdsourcing approach. In total, the CREMA-D corpus contains 7,442 videos with the following distribution: 1,230 happy clips, 1,180 fear clips, 1,222 disgust clips, 1,067 angry clips, 672 sad clips, and 2,071 neutral clips. Our recognition task consists of predicting the six emotional states included in the corpus.

4.2. Visual Features

The visual features are obtained by using the OpenFace toolkit [24] to detect and extract faces from every clip frame from the CREMA-D corpora by estimating bounding boxes using the *multi-task cascaded convolutional neural network* (MTCNN) face detection algorithm [25]. After extracting the bounding boxes, we mask the face area isolating it from the background region. We normalize the pixel intensities between -1 and 1. The images are then resized to the predetermined dimension of $224 \times 224 \times 3$. The normalized image is processed by the VGG-face model [26], which we fine-tune for an emotion recognition task using the AffecNet [27] corpus as a seven-class problem: neutral, happiness, sadness, surprise, fear, disgust, and anger. We attach three *fully connected* (FC) layers to the original VGG-face model and fine-tune this model for 50 epochs using the *adaptive moment estimation* (ADAM) optimizer, with a learning rate set to 0.000075. The facial feature representations at the frame level are obtained from the first fully connected layer of the fine-tuned VGG-Face model, which has a dimension of 4,096. The representations are then concatenated row-wise with all the other frames within each clip from the dataset to be used as input to the video branch of our audiovisual model (i.e., a $N_v \times 4,096$, where N_v is the number of visual frames in the video).

4.3. Acoustic Features

The acoustic features correspond to the *low-level descriptors* (LLDs) included in the feature set of the paralinguistic challenge at Interspeech 2013 [28]. This set is obtained using the OpenSmile toolkit [29]. The LLDs include spectral, prosodic, and energy-based acoustic features extracted at the frame level. The set includes energy, fundamental frequency (f0), and *Mel-frequency cepstral coefficients* (MFCCs). These features have been proven to contain relevant information for SER tasks [17]. The features were extracted using a window length of 32ms with a step size of 16ms. The resulting LLDs consist of 130 frame-based acoustic features, which are Z-normalized. Similar to the video features, the acoustic features are

Table 1. Comparison of our proposed method with baselines. The table reports the average F1, precision, and recall score values across 20 experiments using different random seeds (✶ indicates that our model is significantly better than the other three methods).

CREMA-D						
	Macro			Micro		
Architecture	Prec.	Rec.	F1	Prec.	Rec.	F1
Our Method ✶	**80.3**	**80.4**	**80.2**	**80.3**	**80.3**	**80.3**
Baseline 1 [4]	76.5	75.7	75.5	75.7	75.7	75.5
Baseline 2 [10]	71.6	71.0	70.6	71.0	71.0	71.0
Baseline 3 [3]	60.6	57.8	56.3	58.0	58.0	58.0

concatenated row-wise creating a $N_a \times 130$ input matrix, where N_a is the length of the acoustic sequence.

5. EXPERIMENTAL RESULTS

All the transformer blocks included in the backbone framework are five layers deep, each of them implemented with 10 attention heads. We set a dropout rate of $p = 0.25$ to the output embeddings obtained from the attention layers. The model uses ADAM as the optimizer with an initial learning rate set to 7.25E-04. During training, we have a learning decay set to six epochs. We set the gradient clipping threshold to 0.8 and a batch size of 64. We use the *rectified linear unit* (ReLU) as the activation function. The model was trained for 40 epochs with a model-saving criterion based on improvements in the development loss. To add noise to the noisy encoding layers, we use a variance set to σ^2=0.2. For the ladder network, the hyper-parameter for the reconstruction loss is set to $\lambda_0 = 0.5$, $\lambda_1 = 0.5$, and $\lambda_2 = 0.1$ (Eq. 4). A preliminary search on the development set showed that assigning less weight to the top layer (λ_2) yielded better results. We use the mean squared error as the reconstruction cost and cross-entropy loss for the classification task. The model has $\sim$8.2M parameters, it was coded in Pytorch, and trained using an Nvidia QUADRO RTX 8000.

We compare our model with three baselines to verify our model's performance against strong frameworks. Baseline 1 is the AuxFormer framework proposed by Goncalves and Busso [4, 5]. Baseline 2 is derived from the method presented in Tsai et al. [10], which proposed a multimodal transformer architecture. We implement their model making a few changes to their original architecture to adapt their method from three modalities (text, audio, and video) to two modalities (audio and video). Baseline 3 is implemented using the transformer-based approach proposed by Parthasarathy and Sundaram [3]. Overall, our model has 1.17 times more parameters than [4], 1.24 times more parameters than [10], and 3.7 times more parameters than [3]. All models are trained for 40 epochs.

We train the proposed and baseline models 20 times with different train, development, and test partitions. For each trial, we generate random splits in a speaker-independent manner using 70% of the data for the train set, 15% of the data for the development set, and 15% of the data for the test set. The random partitions are consistently used to train and evaluate each of the models. We evaluated the models' performances using both the macro-averaged and micro-averaged precision, recall, and F1 scores averaged across the 20 trials for each model. Experimental results are recorded and compared using a one-tailed matched pair t-test over the 20 results with a significance level at p-value $<$0.05 to assert statistical significance.

5.1. Comparison with Baselines

We compare our proposed framework's performance with the three popular multimodal emotion recognition baselines. Table 1 shows the precision, recall, and F1 macro and micro scores of the models for the six-class problem on the CREMA-D corpus. Our proposed framework achieves the best scores in all cases, significantly outperforming all the baselines. The results support our hypothesis that using a method that incorporates both forward and backward techniques for representation learning is beneficial in audiovisual settings. Baseline 1 is the AuxFormer architecture [4, 5], which contains mostly the same architecture as our backbone network. By comparing our model's performance with baseline 1 [4], we can quantify the benefits of using the multimodal ladder networks. The use of the proposed auxiliary unsupervised task plays an important role in the performance of our model. This result supports the idea that adding relevant unsupervised auxiliary tasks helps our model capture or learn important audiovisual representations to improve performance on our downstream task.

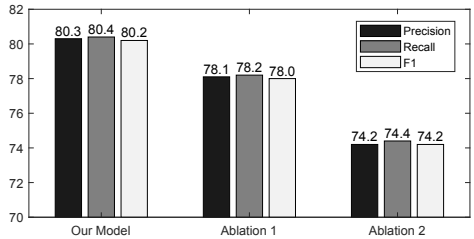

Fig. 3. Ablation analysis reporting macro precision, recall, and F1-score for our proposed frameworks and the ablated versions of our model. Ablation 1 removes the unimodal cross-layer ladder networks, and ablation 2 removes the audiovisual ladder network.

5.2. Framework Evaluations

We perform ablations to our framework by removing part of the proposed multimodal ladder networks. We compare these performances with the one obtained by our complete framework following the same training settings. We perform two ablations to our framework. Ablation 1 consists of removing the unimodal cross-layer ladder network predictions. This block is depicted inside the blue dashed line in Figure 1. Ablation 2 consists of removing the audiovisual ladder network depicted in Figure 1 in the orange box. Figure 5.2 shows the results. There is a performance drop for both ablations, showing that both components are needed. Ablation 1 leads to a drop of 2.2% in performance. Ablation 2 results in an even larger drop, reducing the performance metrics up to 6.1% (absolute) compared to results achieved by our full framework. Both ladder network mechanisms are important for the overall performance of the model.

6. CONCLUSIONS & FUTURE WORK

This work proposed an audiovisual representation learning framework based on a multimodal ladder network, which uses the reconstruction of intermediate hidden representations across modalities. The proposed approach achieves state-of-the-art performance on a six-class problem on the CREMA-D corpus (i.e., emotional categories). Our proposed model utilizes forward and backward learning techniques to build representations that capture both cross-modal and modality-specific information for our downstream task. Our results show that the learned representations carry meaningful information providing significantly better performance than the baselines.

This framework can be expanded to explore the potential of using large amounts of unlabeled data, implementing the multimodal ladder networks using a semi-supervised setting. This strategy provides the opportunity to train the model with labeled (source domain) and unlabeled (target domain) data, enhancing the audiovisual representations. The strong results obtained in this study suggest that exploring extensions of the proposed model can lead to important improvements in multimodal emotion recognition.

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] 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.

[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] L. Goncalves and C. Busso, "AuxFormer: Robust approach to audiovisual emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, May 2022, pp. 7357–7361.

[5] L. Goncalves and C. Busso, "Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2156–2170, October-December 2022.

[6] T. Baltrušaitis, C. Ahuja, and L. P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, February 2019.

[7] C. Busso and S.S. Narayanan, "Interplay between linguistic and affective goals in facial expression during emotional utterances," in *7th International Seminar on Speech Production (ISSP 2006)*, Ubatuba-SP, Brazil, Dec. 2006, pp. 549–556.

[8] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *AAAI Conference on Artificial Intelligence (AAAI 2021)*, Virtual Conference, February 2021, vol. 35, pp. 10790–10797.

[9] A. Zadeh, P.P. Liang, J. Vanbriesen, S. Poria, E. Tong, E. Cambria, M. Chen, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *ACM Association for Computational Linguistics (ACL 2004)*, Melbourne, Australia, July 2018, vol. 1, pp. 2236–2246.

[10] 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.

[11] 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.

[12] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and -specific representations for multimodal sentiment analysis," in *ACM International Conference on Multimedia (MM 2020)*, Seattle, WA, USA, October 2020, pp. 1122–1131.

[13] 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.

[14] S. Parthasarathy and C. Busso, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 3698–3702.

[15] J.-H. Tao, J. Huang, Y. Li, Z. Lian, and M.-Y. Niu, "Semi-supervised ladder networks for speech emotion recognition," *International Journal of Automation and Computing*, vol. 16, no. 4, pp. 437–448, August 2019.

[16] S.-G. Leem, D. Fulford, J.-P. Onnela, D. Gard, and C. Busso, "Separation of emotional and reconstruction embeddings on ladder network to improve speech emotion recognition robustness in noisy conditions," in *Interspeech 2021*, Brno, Czech Republic, August-September 2021, pp. 2871–2875.

[17] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2697–2709, September 2020.

[18] J. Huang, Y. Li, J. Tao, Z. Lian, M. Niu, and J. Yi, "Speech emotion recognition using semi-supervised learning with ladder networks," in *Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia 2018)*, Beijing, China, May 2018, pp. 1–5.

[19] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Association for Computational Linguistics (ACL 2018)*, Melbourne, Australia, July 2018, pp. 2247–2256.

[20] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, Copenhagen, Denmark, September 2017, pp. 1103–1114.

[21] 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.

[22] 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.

[23] M. Pezeshki, L. Fan, P. Brakel, A. Courville, and Y. Bengio, "Deconstructing the ladder network architecture," in *International Conference on Machine Learning (ICML 2016)*, New York, NY, USA, June 2016, pp. 2368–2376.

[24] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *IEEE Conference on Automatic Face and Gesture Recognition (FG 2018)*, Xi'an, China, May 2018, pp. 59–66.

[25] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, October 2016.

[26] 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.

[27] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, January-March 2019.

[28] B. Schuller et al., "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.

[29] 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.