IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 19, 2022, accepted June 7, 2022, date of publication June 16, 2022, date of current version June 22, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3183587

Can We Exploit All Datasets? Multimodal Emotion Recognition Using Cross-Modal Translation

YEO CHAN YOON

Department of Artificial Intelligence, Jeju National University, Jeju 63243, South Korea

e-mail: ycyoon@jejunu.ac.kr

This work was supported in part by the Ministry of Science and ICT (MSIT), South Korea, under the National Program for Excellence in Software (SW) supervised by the Institute of Information and Communications Technology Planning and Evaluation (IITP), under Grant 2018-0-01863; and in part by the Ministry of Culture, Sports and Tourism and Korea Creative Content Agency under Project R2021040097.

ABSTRACT The use of sufficiently large datasets is important for most deep learning tasks, and emotion recognition tasks are no exception. Multimodal emotion recognition is the task of considering multiple types of modalities simultaneously to improve accuracy and robustness, typically utilizing three modalities: visual, audio, and text. Similar to other deep learning tasks, large datasets are required. Various heterogeneous datasets exist, including unimodal datasets constructed for traditional unimodal recognition and bimodal or trimodal datasets for multi-modal emotion recognition. A trimodal emotion recognition model shows high performance and robustness by comprehensively considering multiple modalities. However, the use of unimodal or bimodal datasets in this case is problematic. In this study, we propose a novel method to improve the performance of emotion recognition based on a cross-modal translator that can translate between the three modalities. The proposed method can train a multimodal model based on three modalities with different types of heterogeneous datasets, and the dataset does not require alignment between modalities: visual, audio, and text. We achieved a high performance exceeding the baseline in CMU-MOSEI and IEMOCAP, which are representative multimodal datasets, by adding unimodal and bimodal datasets to the trimodal dataset.

INDEX TERMS Deep learning, emotion recognition, generative adversarial networks, machine learning, multimodal emotion recognition.

I. INTRODUCTION

The perception of human emotions is becoming an essential part of various human-computer interaction systems, as recognizing human emotions affect plays a crucial role in our daily lives. People respond to and act according to their perceptions of emotions in response to external stimuli. Intelligent systems, such as surveillance, robotics, and medical systems, benefit from the ability of understanding human emotions and behaviors.

One of the most important tasks in recognizing emotions is to assemble various types of information that express human emotions. The expression of human emotions is intrinsically multi-modal. Voice pitch, speed, facial expression, words used, and gestures are among several means of expressing emotions. Therefore, intuitively, using various modalities can achieve higher performance and reliability compared to a limiting use of one modality. One of the main challenges in multi-modal emotion recognition is the difficulty in obtaining labeled data because it takes a long time for humans to identify categories of emotions in video, audio, or text. Owing to the efforts of several researchers to recognize emotions, labeled image-based facial expression recognition datasets [1]–[3] or text-based emotion recognition datasets [4]–[6] became publicly available. However, the unified labeling set for video, audio, and text modalities is much smaller than that for single or bimodal datasets. Building large-scale multimodal datasets for video, audio, and text is expensive and time-consuming. The main motivation of this study is to investigate the effective utilization of datasets with different modal information by using a learning strategy to train tri-modal emotion recognition with cross-modal translators.

To utilize datasets with different modalities, many researchers have proposed cross-modal transferring methods; target modal data are augmented through cross-modal translation with source modal data and used to train a target-modal-based recognition model. For example, to transfer visual information to audio, He *et al.* [7] used VAEGAN [8] as a visual-to-audio translator. The conditional generative adversarial network(GAN) [9] and cycle GAN [10] have also been

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wei.

used [7], [11], [12] to translate visual to audio information. For audio-text transfer, a consistent prediction method for real speech and synthetic speech has been proposed [13] to improve the speech recognition performance. Yoon *et al.* [14] translated birds and plants images into text to accurately classify birds and plants.

In this study, we propose a multi-modal emotion recognition model that takes the three modalities of video, audio, and text as input and learns from a multimodal dataset containing all three modalities as well as single or bimodal datasets. To this end, we propose a feature-level cross-modal transfer model for translation between the three modalities. Data, including video, audio, and text used for emotion recognition, were expressed in a time series. Therefore, to transfer a word into an audio signal, word-level multimodal alignment is required. However, aligning different modalities generally requires human labor. To address this problem, we propose a novel cross-modal translation model using a sequence-level discriminator for unaligned multimodal datasets. Using additional heterogeneous single- or bimodal datasets, we prove that the proposed method is effective in improving performance.

We trained a cross-modal translator and a multimodal emotion recognizer with an end-to-end architecture that simultaneously learns both models. We tested the performance of the proposed end-to-end cross-modal translation and emotion recognition model by applying it to two benchmark datasets, CMU-MOSEI and IEMOCAP. The contributions of this study are as follows.

1) We proposed a strategy for training a multimodal emotion recognition model using multiple heterogeneous datasets with different modalities. We used cross-modal translators and an end-to-end learning strategy to achieve this goal. Cross-modal translators can be used to leverage single or bimodal datasets and improve the performance of emotion recognizers based on data augmentation effects.
2) We propose a novel cross-modal translation model between trimodal unaligned multimodal datasets. By adding a sequence-level discriminator, we can train a cross-modal translator without a human-labored word or phoneme-level alignment job. To the best of our knowledge, this is the first attempt at recognizing emotion by augmenting three modals: visual, audio, and text.
3) Our approach is evaluated on the representative multimodal emotion recognition benchmark datasets CMU-MOSEI and IEMOCAP; it exceeds the baseline approach by 13.4% on CMU-MOSEI and 10.4% on IEMOCAP.

II. RELATED WORKS

A. MULTIMODAL EMOTION RECOGNITION

Many prior studies have been conducted on multi-modal emotion recognition. In recent years, considerable progress has been made in this area by using modality fusion methods. A dynamic fusion graph-based network [15], which fuses modalities dynamically in a hierarchical manner, a tensor fusion network that combines data representation from each modality to an embedding [16], a capsule GCN considering information redundancy and complementarity [17], and late or early fusion networks [15], [18]–[20], which emphasizes a relative place of network have been proposed and showed better performance than single modality emotion recognition systems. M3ER [21] also uses multiplicative fusion to determine a more important modality on a per-sample basis.

Previous fusion methods generally do not require alignment between modalities, and they are difficult to interact in an intermodal sequential manner. To overcome this limitation, the attention-mechanism-based fusion method [22]–[24] or transformer algorithm [25]–[28] are widely used. The transformer [29] algorithm uses self-attention to analyze the correlation between items constituting a sequence. Tsai *et al.* [25] introduced a cross-modal transformer. They proposed a multimodal transformer that provides latent cross-modal adaptation, which fuses multimodal information by directly attending to low-level features in other modalities.

To improve the performance with additional datasets, a self-supervised training strategy that uses a pre-training method with a large-scale unlabeled dataset is used for emotion recognition tasks [26]–[28]. Rahman *et al.* [26] deployed BERT [30] and XLNet [31] with multi-modal adaptation gates. Khare *et al.* [28] trained a transformer on a masked language-modeling task for trimodal emotion recognition. They used a cross-modal-based transformer model to analyze the input modalities in an intermodal sequential manner.

In this study, we also used a cross-modal based multimodal transformer. However, to compensate for insufficient data, we deploy a data augmentation method instead of fine-tuning pre-trained transformers that require large-scale computing resources.

B. CROSS-MODAL TRANSLATION

To address the problem of data shortage and imbalance between modalities, recent studies on data augmentation through cross-modality translation have been conducted. Projecting different modalities onto a shared semantic space is a commonly used method for representing and manipulating multiple modalities. Harwath *et al.* [32] proposed a method for projecting audio and images onto a shared embedding space and clustering embedding to translate them into related text words. This method allows reading the text searches using only the image and audio information. Qi *et al.* [33] also used a shared semantic space for image-text translation. They treated images and texts in two different languages, trained a cross-modal translation model using reinforcement learning, and then applied the training results to a cross-modal retrieval task. To analyze sentiments, Yang *et al.* [34] proposed a method for visual-to-text and audio-to-text translation with a pre-trained BERT using a shared semantic space.

While a simple linear mapping function for cross-modal translation is used for the fusion based multi-modal emotion recognition method [16], [21], standalone generative models such as GANs or autoencoders can be deployed for cross-modal translation [7], [11], [12], [33]–[35]. Tsai *et al.* [35] introduced an autoencoder-based modality reconstruction method for the missing modalities. To augment audio datasets for audio emotion recognition, He *et al.* [7] introduced a visual-to-audio translator based on VAEGAN with cycle reconstruction loss.

Although prior studies have made progress in this task, few attempts have been made to translate all three modalities and use them to train an integrated multimodal classifier. Additionally, little effort has been made into using multiple heterogeneous datasets together, and there has been no attempt to train an emotion recognizer and cross-modal translator simultaneously in an end-to-end manner, other than by augmenting and feeding the data in a pipeline manner. In this study, we propose a method to address data shortage and imbalance problems by simultaneously training a generative model and a classification model for visual, audio, and text trimodal.

III. METHODOLOGY

We denote the audio, visual, and text modalities as $m \in a, v, t$. We embed the modalities into a shared latent semantic space and denote them as F_a, F_v, F_t. To feed missing modalities into the trimodal emotion recognition model, we translate the given modality inputs into missing modalities with the translators $T_{m1 \to m2}$ where m_1 and $m_2 \in \{a, v, t\}$.

A. END2END CROSS-MODAL TRANSLATION AND EMOTION RECOGNITION ARCHITECTURE

Fig. 1 shows the architecture of the proposed multimodal emotion recognition system using the cross-modal translation method. For the trimodal data, we use the feature extractor FE_m to extract the feature F_m from each modal and feed it to the corresponding transformer module. The final output P_{avt} is the result of the softmax and feedforward layer with the Multimodal Fusion module, which takes the weighted sum of each modality module output as input. Equations (1–3) shows how the final output P_{avt} can be derived from the modality modules MM_m and fusion module MFM. We set a fixed weight of 0.33 for w_a, w_v and w_t for the experiments.

$$WS_{avt} = \sum\nolimits_{m \in a,v,t} w_m MM_m\left(F_m\right) \tag{1}$$

$$\boldsymbol{MF_{avt}} = \boldsymbol{MFM}\left(\boldsymbol{WS_{avt}}\right) \tag{2}$$

$$\boldsymbol{P_{avt}} = \boldsymbol{Softmax}\left(\boldsymbol{FF}\left(\boldsymbol{MF_{avt}}\right)\right) \tag{3}$$

One crucial factor that should be considered in sequential data is the focus on important clues. For example, if one focuses on a moment with a strong facial expression, a clear emotional vocabulary, or a strong tone of voice, emotions can be more easily recognized. A transformer is a well- known neural network model that best reflects this characteristic. Positional embedding [29] were added to the input features to account for the order of sequence components. We also used a transformer for the multimodal fusion module to learn how to combine the results of each modality.

When a unimodal or bimodal sequence is provided, the missing modality is augmented using a cross-modal translator and fed as an input to the corresponding transformer module. In Fig. 1, the visual feature F_v is translated into the audio feature F_a' and the text feature F_t', via a visual-to-audio and visual-to-text translator; then, they are fed into the corresponding transformer module. Equations (4–8) shows how the final output P_{avt} can be derived from the cross-modal translator $T_{v \to a}$, $T_{v \to t}$,modality modules MM_m and fusion module MFM.

$$F_a' = T_{v \to a}\left(F_v\right) \tag{4}$$

$$\boldsymbol{F_t'} = \boldsymbol{T_{v \to t}}\left(\boldsymbol{F_t}\right) \tag{5}$$

$$\boldsymbol{WS_{avt}} = \boldsymbol{w_a MM_a}\left(\boldsymbol{F_a'}\right) + \boldsymbol{w_v MM_v}\left(\boldsymbol{F_v}\right) + \boldsymbol{w_t MM_t}\left(\boldsymbol{F_t'}\right) \tag{6}$$

$$\boldsymbol{MF_{avt}} = \boldsymbol{MFM}\left(\boldsymbol{WS_{avt}}\right) \tag{7}$$

$$\boldsymbol{P_{avt}} = \boldsymbol{Softmax}\left(\boldsymbol{FF}\left(\boldsymbol{MF_{avt}}\right)\right) \tag{8}$$

B. CROSS-MODAL TRANSLATOR FOR SEQUENTIAL INPUT

He *et al.* [7] proposed the VAEGAN-based visual-to-audio modal translator. They augmented audio emotion data with a translator. However, the previous work managed only a single image and audio spectrum. In this study, we propose a sequential VAEGAN for trimodalities, visual, audio, and text. As we use three modalities and pair each modality, our proposed translation unit has six VAEGANS and each VAEGAN has a feature extractor *FE*, an encoder *Enc*, a Decoder *Dec*, and a discriminator *Dis* for a single image slide, a word, and an audio piece. The translator also includes the sequence discriminator *SeqDis* for the entire input sequence. Fig. 2 shows the visual-to-text translator. The visual sequences fed into the visual feature extractor and fake text feature sequence can be generated along with the visual encoder and text decoder with the discriminators. Each Enc_m, Dec_m and Dis_m processes one slide at a time, not whole sequences at once, whereas $SeqDis_m$ manages sequential input to classify the input feature more accurately. $SeqDis_m$ uses a (K + 1)-class objective; K classes are used for ground-truth samples, and the (K + 1)-th class is used for fake samples. We used a transformer-based sequential classifier for the discriminator.

The training objective includes four components: the VAE, GAN, sequential discriminator, and cycle losses.

$$\begin{aligned} \boldsymbol{SL_{VAE}} = &\sum\nolimits_{\boldsymbol{i=1}}^{|\boldsymbol{m}|} \sum\nolimits_{\boldsymbol{j=1}}^{|\boldsymbol{x_{m_i}}|} \boldsymbol{L_{VAE_{m_i}}} \\ &\times \left(\boldsymbol{Dis_{m_i}}\left(\boldsymbol{x^j_{m_i}}, \boldsymbol{Dec_{m_i}}\left(\boldsymbol{Enc_{m_i}}\left(\boldsymbol{x^j_{m_i}}\right)\right)\right)\right) \end{aligned} \tag{9}$$

$$\begin{aligned} \boldsymbol{SL_{GAN}} = &\sum\nolimits_{\boldsymbol{i=1}}^{|\boldsymbol{m}|} \sum\nolimits_{\boldsymbol{j \neq i}}^{|\boldsymbol{m}|} \sum\nolimits_{\boldsymbol{k=1}}^{|\boldsymbol{x_{m_i}}|} \boldsymbol{L_{GAN_{m_i \to m_j}}} \\ &\times \left(\boldsymbol{Dis_{m_j}}\left(\boldsymbol{GT^k_{m_j}}, \boldsymbol{Dec_{m_j}}\left(\boldsymbol{Enc_{m_i}}\left(\boldsymbol{x^k_{m_i}}\right)\right)\right)\right) \end{aligned} \tag{10}$$

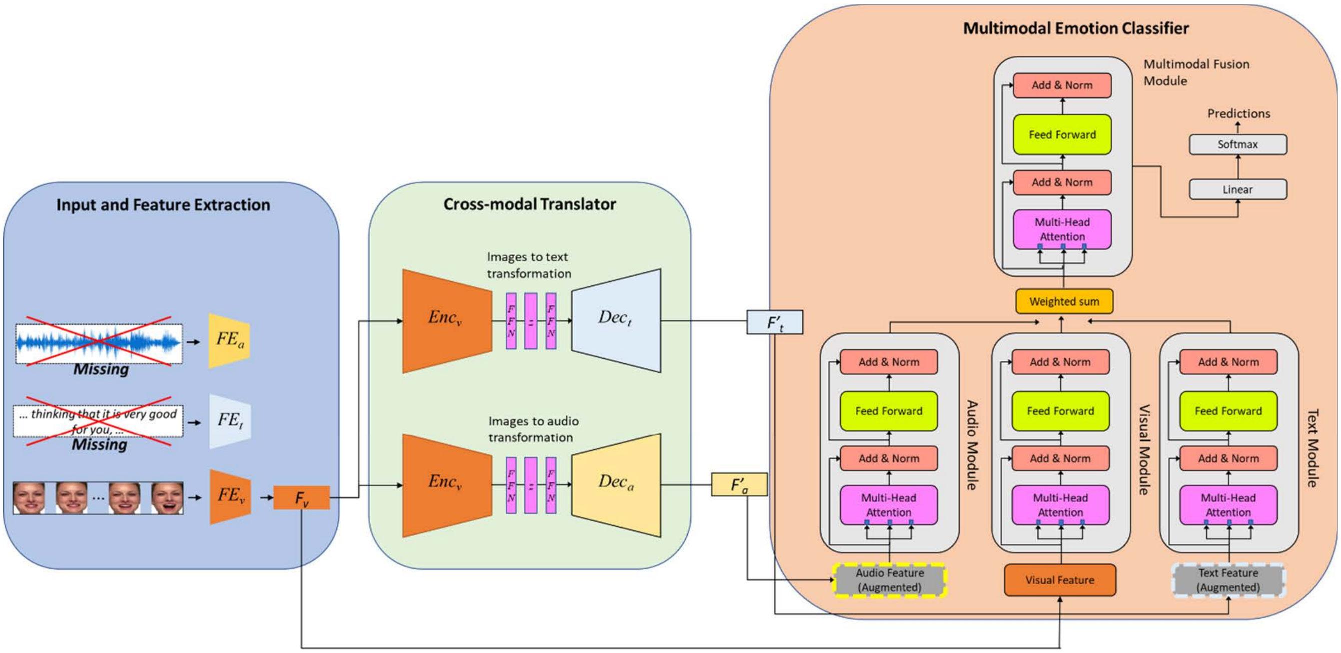

FIGURE 1. Architecture of proposed Multimodal Emotion Classifier with Cross-modal Translator. Missing modalies can be augmented using the Cross-modal Translator and fed into the Tri-modal Multimodal Emotion Classifier.

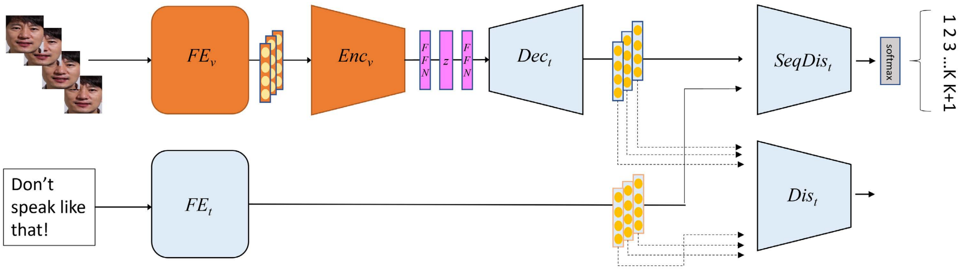

FIGURE 2. Architecture of proposed Cross-modal Translator. The sequence discriminator *SeqDis* manages sequential input to classify the input feature more accurately wheres the Discriminator *Dis* processes one slide at a time, not whole sequences at once.

$$SL_{SeqDis} = \sum_{i=1}^{|m|}\sum_{j\neq i}^{|m|} L_{SeqDis_{m_i}} \times \left(x_{m_j}, Dec_{m_j}\left(Enc_{m_i}\left(x_{m_i}\right)\right), y\right) \tag{11}$$

$$SL_{Cycle} = \sum_{i=1}^{|m|}\sum_{j\neq i}^{|m|}\sum_{k=1}^{|x_{m_i}|} L_{Cycle_{m_i\to m_j\to m_i}} \times \Big(Dis_{m_j}\Big(x^k_{m_i}, Dec_{m_i}\Big(Enc_{m_j} \times \Big(Dec_{m_j}\Big(Enc_{m_i}\Big(x^k_{m_i}\Big)\Big)\Big)\Big)\Big)\Big) \tag{12}$$

$$\min_{(Enc_m, Dec_m)} \max_{(Dec_m)} \underset{(m\in a,v,t)}{E} \times \left(SL_{VAE} + SL_{GAN} + SL_{SeqDis} + SL_{Cycle}\right) \tag{13}$$

$x^j_{m_i}$ indicates j-th constituent of sequence x of i-th modality. L_{VAE} indicates variational auto encoder loss for input modal m_i · $GAN_{m_i\to m_j}$ converts input model m_i to output modal m_j and $L_{GAN_{m_i\to m_j}}$ indicates the loss of the GAN. The cycle reconstruction loss L_{cycle} is also employed to reflect the two-direction translation $m_i \to m_j$ and $m_j \to m_i$. The objective function of $SeqDis_m$ can be calculated as follows:

$$\max_{SeqDis_{m_j}} E_{x\in D_g} logP_{SeqDis_{m_j}}\left(k+1|x^1, x^2, \ldots, x^n\right) + E_{x\in D_{gt}} P_{SeqDis_{m_j}}\left(y|x^1, x^2, \ldots, x^n\right) \tag{14}$$

x^i indicates i-th constituent of sequence x, D_g indicates the generated sequence, and D_{gt} indicates the ground-truth examples. To calculate $L_{GAN_{m_i\to m_j}}$, aligned multi-modal ground truth data were required. If the given data have no paired aligned ground truth data, we update the modules, L_{VAE}, L_{cycle}, and L_{SeqDis} except L_{GAN}.

C. TRAINING STRATEGY

The proposed model uses an end2end strategy that simultaneously learns the cross-modal translator and the multimodal emotion classifier. To improve the learning performance with

Algorithm 1

```
for data in dataset:
 audio, visual, text = data.input # given modality inputs
 augmented_inputs =[]
 if audio != None:
     audios.append(audio)
     translator_loss, fake_visual =
                                        CMTranslatorAtoV
                                            (data.input)
     visuals.append(fake_visual)
     translator_loss.backward()
     translator_loss, fake_text =
                                        CMTranslatorAtoT
                                            (data.input)
   texts.append(fake_text)
   translator_loss.backward()
if visual != None:
   visuals.append(visual)
   translator_loss, fake_audio =
                                        CMTranslatorVtoA
                                            (data.input)
   audios.append(fake_audio)
   translator_loss.backward()
   translator_loss, fake_text =
                                        CMTranslatorVtoT
                                            (data.input)
   texts.append(fake_text)
   translator_loss.backward()
if text != None:
   texts.append(text)
   translator_loss, fake_text =
                                        CMTranslatorTtoA
                                            (data.input)
   texts.append(fake_text)
   translator_loss.backward()
   translator_loss, fake_visual =
                                        CMTranslatorTtoV
                                            (data.input)
   visuals.append(fake_visual)
   translator_loss.backward()
for audio in audios:
   for visual in visuals:
      for text in texts:
         loss, pred = MultimodalTransformer(audio, visual,
                                          text)
         loss.backward()
```

cross-modal data augmentation, possible fake features were generated from real examples. Algorithm 1 describes the learning strategy in detail.

Algorithm 1. Training strategy with cross-modal translator and multimodal emotion classifier.

IV. EXPERIMENTS

A. DATASETS

To evaluate the performance of the emotion recognition task, we applied the proposed method to the CMU Multimodal Opinion sentiment and emotion intensity (CMU-MOSEI) dataset [38] and the interactive emotional dyadic motion capture database (IEMOCAP) dataset [39]. The CMU-MOSEI dataset is currently the largest publicly available multi-modal dataset for emotion recognition. It comprises 23,453 single-speaker video segments. 1,000 distinct speakers and 250 topics were acquired from YouTube. The dataset consists of six emotions: happiness, sadness, anger, surprise, fear, and disgust. In addition to the visual and audio data, human-labeled transcriptions were included for linguistic emotion analysis. Detailed statistics are presented in Table 1.

TABLE 1. Statistics of CMU-MOSEI dataset.

	Train	Valid	Test
Happy	8147	13131	2522
Sad	3906	576	1334
Anger	3443	427	971
Surprise	1562	201	479
Disgust	2720	352	922
Fear	1319	186	332

The IEMOCAP dataset was built for multimodal human emotion analysis. It was recorded from ten actors in dyadic sessions with markers on the face that provided detailed information about their facial expressions during scripted and spontaneous spoken communication scenarios. It contains four labelled emotion annotations: angry, happy, neutral, and sad. Detailed statistics are presented in Table 2.

To gain accuracy from a cross-modal translator, we utilized single-or bimodality emotion recognition datasets: AFEW [40] (video and audio), CK+ (video) [3], RAVDESS (video and audio) [41], and SemEval 2018 E-c (text) [4]. The AFEW contains videos from different movies and TV series with spontaneous expressions. The training, validation, and test sets contained 773, 383, and 653 video files, respectively. CK+ consisted of 529 videos from 123 subjects, ranging from 18 to 50 years old of age, with a variety of genders. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 7,356 files with 24 distinct professional actors (12 females and 12 males). SemEval 2018 E-c (SemEval), a multilabeled text emotion dataset, comprises 10,983 tweets and 11 labels for the presence or absence of emotions.

For these datasets, we only used data that shared emotions with CMU-MOSEI, and these datasets were merged into a larger dataset.

B. TRAINING DETAILS

The model was trained using the Adam optimizer with a learning rate 0.01 and 64 batch. The detailed settings for the feature extractor, transformer-based emotion recognizer, and cross-modal translator are as follows.

1) FEATURE EXTRACTION

To extract features from the visual, audio, and text data, we deployed different feature extractors for each modality. For the visual modality, we deployed the video-based facial expression recognizer, FAN [42]. This method uses frame attention to automatically highlight discriminative frames from input video. For audio feature extraction, we follow

TABLE 2. Statistics of IEMOCAP dataset.

	Train	Valid	Test
Happy	338	116	135
Sad	690	188	193
Anger	735	136	227
Natural	954	358	383

the method and setting of the audio classification algorithm, Panns [43], with the learnable audio frontend (LEAF) [44] instead of Mel-filterbanks. For the visual and audio feature extractor, we added a fully connected layer with 300 hidden units to feed the output to the input of the cross-modal translator and multi-model emotion recognizer. For text modality, glove word embedding [45] was used to extract word vectors from the transcripts. We deployed the setting of the CMU-MOSEI SDK [46] to learn the embeddings.

2) MULTI-MODAL EMOTION RECOGNITION

Each transformer for the audio, video, text, and multimodal fusion modules was configured with the same model architecture. The model had a feed-forward layer of dimension 128 and four attention heads. The number of hidden nodes of attention was 128.

3) CROSS-MODAL TRANSLATION MODULE

Based on the work in [47], we added the *SeqDis* discriminator for sequential features. *SeqDis* consists of one transformer encoder with a feed-forward layer of dimension 128 and four attention heads. The number of hidden nodes of attention was 128.

We shared the weights of the last layer of the encoder and decoders for each modality translator to embed features on the same latent semantic spaces.

C. RESULTS

We used the weighted accuracy (WA) [48] and F1-score for each emotion owing to natural imbalances across various emotions. Table 3 lists the performance of the proposed models on the CMU-MOSEI dataset. For comparison, we included the graph memory fusion network (GraphMFN) [38], which was published along with the CMU-MOSEI dataset, and Khare's cross-modal transformer-based multimodal emotion recognition method [28]. In Table 3, the Multimodal transformer indicates the proposed transformer-based multimodal emotion recognition model without a cross-modal translation module. +Cross-modal translation shows the performance when applying the data augmentation strategy in Algorithm 1 using the proposed cross-modal translation model and a multimodal transformer. Through data augmentation, the performance improved by an average of 1.6% in terms of WA and 2.7% on average in terms of F1 measure. When the proposed *SeqDist* was used, the performance was improved by 2.2% based on WA and 3.0% based on F1. + The auxiliary dataset indicates the model performance when unimodal and bimodal datasets are added, in addition to the CMU-MOSEI datasets. A significant improvement of 5.4% in terms of WA and 8.5% in terms of the F1 measure was observed. A similar trend was observed in the additionally tested IEMOCAP dataset. When the proposed method and additional datasets were used, the performance improved by 10.4% in the WA standards and 14.8% in the F1 score, demonstrating the effectiveness of the proposed method.

1) CONFUSION MATRIX

We show the per-class performance of the proposed model with an auxiliary dataset on CMU-MOSEI and IEMOCAP using the confusion matrix in Fig. 3. Using the auxiliary dataset, the proposed algorithm achieved an accuracy of over 70% for each class. Owing to data imbalance, data samples tend to be misclassified into classes with more samples. For example, the most common type of error is the misclassification of samples into happy classes in CMU-MOSEI and neutral classes in IEMOCAP.

2) EFFECT OF DATA AUGMENTATION

Tables 5 and 6 show the results of analyzing the data augmentation effect. When data are augmented through cross-modal translation, even when using only 90% of the total data, the performance is better compared to learning with a 100% dataset without data augmentation in F1 measurements. In addition, using data augmentation improves the existing performance regardless of the data size. However, the larger the data size, the better the growth effect. It appears that the augmentation effect increases as the number of training data increases because a sufficient amount of training data is required for cross-modal translation learning.

3) ABLATION STUDY

To determine how much individual modality affects the model, we conducted an ablation experiment. Tables 7 and 8 show the performance changes when the data were augmented for each modality. T, V, and A represent the text, video, and audio data, respectively. In both datasets, the greatest performance improvement was achieved when augmentation was applied to all the modalities. In particular, the IEMOCAP dataset exhibited the highest performance for all the classes.

4) EFFECT OF END2END STRATEGY

Table 9 compares the performances of the pipeline and end2end strategies. In the pipeline strategy, we train the cross-modal translator first and then train the multimodal transformer-based emotion.

In the end2end strategy, we trained the cross-modal translator and multimodal emotion classifier simultaneously; we set the ratio of n emotion classifier iterations per translator update and compare the performances. The results in Table 9 show that, on average, the performance of the end2end strategy with a 1:5 and 1:7 balance is better than

TABLE 3. Model performance comparison on CMU-MOSEI dataset.

Method	Anger		Disgust		Fear		Happy		Sad		Surprise		Average	
	WA	F1	WA	F1	WA	F1	WA	F1	WA	F1	WA	F1	WA	F1
GraphMFN	62.6	72.8	69.1	76.6	62	**89.9**	66.3	66.3	60.4	66.9	53.7	85.5	62.4	76.3
Khare	68.2	74.7	74.8	**82.4**	61.5	86.5	67.4	67.1	64.6	**72.5**	62.9	**88.1**	66.6	**78.5**
Multimodal Transformer	80.7	67.1	79.0	64.0	**87.0**	71.7	74.5	68.4	74.1	60.0	86.9	73.6	80.4	67.4
+Cross-modal translation	83.2	71.2	83.7	68.8	84.4	71.6	78.4	73.6	75.0	61.5	87.0	74.1	82.0	70.1
+ SeqDist	**86.5**	75.7	84.0	69.5	79.0	56.0	80.4	76.5	78.7	66.8	86.9	78.0	82.6	70.4
+ Auxiliary dataset.	84.0	**76.5**	**87.1**	75.2	84.9	70.7	**85.0**	**81.4**	**81.7**	70.9	**92.3**	80.9	**85.8**	75.9

TABLE 4. Model performance comparison on IEMOCAP dataset.

Proposed	Happy		Sad		Angry		Neutral		Average	
	WA	F1	WA	F1	WA	F1	WA	F1	WA	F1
Multimodal transformer	73.8	58.2	74.1	58.1	70.7	57.4	72.8	65.7	72.8	59.9
+Cross-modal translation	77.0	63.8	77.9	64.2	76.0	65.1	77.9	72.3	77.2	66.3
+ SeqDist	76.7	63.2	80.2	68.0	77.2	67.3	79.5	73.9	78.4	68.1
+ Auxiliary dataset.	**80.2**	**68.5**	**84.0**	**73.1**	**84.1**	**75.9**	**84.3**	**81.3**	**83.2**	**74.7**

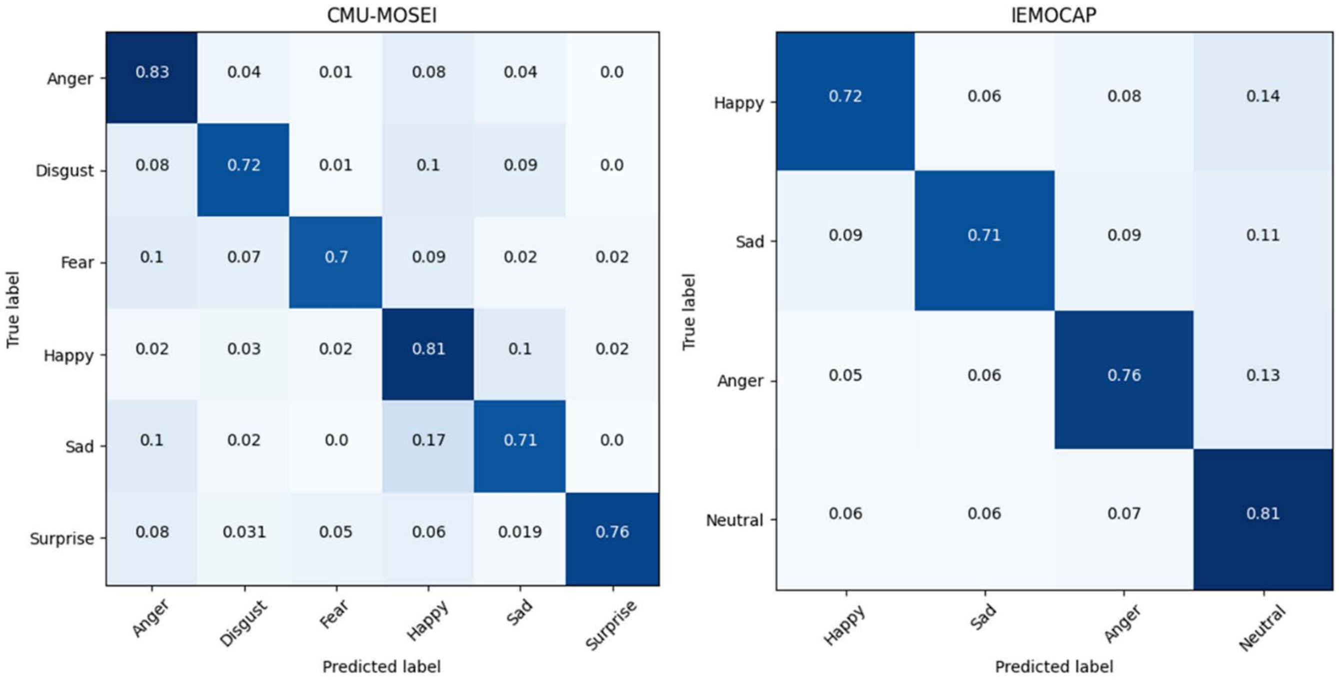

FIGURE 3. Confusion matrix on CMU-MOSEI, IEMOCAP dataset for the proposed method with auxiliary single or double modality dataset.

that of the pipeline strategy. In particular, end2end with a 1:7 balance showed the best overall performance. Whenever a fake feature is generated using the translator and classification is performed using the fake features, both the translator and classifier are updated according to the backpropagation algorithm. Experimental results confirmed that the performance was improved if only the translator was occasionally updated. However, when the rate of learning only the translator was relatively high (1:3), the performance was poor compared to the pipelined method.

TABLE 5. Data size reduction test on CMU-IEMOCAP dataset.

data usage rate	w/o data augmentation		with data augmentation	
	WA	F1	WA	F1
70%	65.3	58.2	66.31	59.4
80%	71.2	60.1	73.12	63.8
90%	75.1	64.2	78.17	70.1
100%	80.4	67.4	**82.6**	**70.4**

TABLE 6. Data size reduction test on IEMOCAP dataset.

data usage rate	w/o data augmentation		with data augmentation	
	WA	F1	WA	F1
70%	60.1	54.7	61.32	56.1
80%	65.1	57.1	67.2	56.8
90%	69.4	56.8	74.15	59.3
100%	72.8	59.9	**78.4**	**68.1**

TABLE 7. Ablation studies on CMU-MOSEI dataset.

Metric	Weighted Accuracy					
Modality	Angry	Disgust	Fear	Happy	Sad	Surprise
T+A+V	84.0	**87.1**	**84.9**	**85.0**	81.7	**92.3**
T+A	83.5	85.3	82.8	83.2	**82.1**	87.2
T+V	81.6	85.1	80.24	82.1	79.1	89.1
A+V	**85.7**	86.5	77.8	84.1	80.8	88.4
T	80.9	84.8	79.2	81.5	77.5	85.1
A	81.2	83.6	76.4	82.1	78.2	87.0
V	84.3	84.2	75.1	82.8	80.2	88.2

TABLE 8. Ablation studies on IEMOCAP dataset.

Metric	Weighted Accuracy			
Modality	Happy	Sad	Angry	Neutral
T+A+V	**80.2**	**84.0**	**84.1**	**84.3**
T+A	78.1	81.3	80.3	82.1
T+V	77.5	82.5	82.4	84.5
A+V	79.3	83.1	83.5	83.2
T	75.2	79.4	80.1	81.7
A	78.0	80.9	81.0	80.8
V	77.9	81.8	82.0	82.8

In fact, it seems that if the translator is frequently trained alone, the features are translated independently of the classifier.

TABLE 9. Comparison of end2end training strategies.

learning ratio	CMU-MOSEI		IEMOCAP	
	WA	F1	WA	F1
pipeline	82.4	72.2	80.1	**76.0**
1:3	80.2	68.5	82.3	72.2
1:5	83.1	72.3	**83.2**	74.7
1:7	**85.8**	**74.7**	81.8	75.4
1:9	82.2	71.5	79.1	73.4

V. CONCLUSION AND FUTURE WORKS

We presented the multimodal emotion recognition model that used cross-modal translators. Using the proposed method, we can further exploit the heterogeneous types of datasets with different modalities.

For inter-modal translation, we proposed novel cross-modal translators that uses a sequential discriminator to cover unaligned multimodal sequence data. The proposed model learns cross-modal translators and multimodal emotion recognizers simultaneously, and this strategy further improves the performance. The empirical results demonstrate that the proposed method is efficient in handling multiple datasets with different modalities. With our method, the cost of constructing, aligning, or reorganizing the dataset can be significantly decreased. In a future work, we shall apply our method to self-supervised learning using multimodal datasets. It is well known that self-supervised learning strategies can help improve robustness and performance. To train a multimodal model, we expect that unlabeled heterogeneous datasets could be helpful, and that cross-modal translators would become more robust in the self-supervised learning process.

REFERENCES

[1] I. Gogić, M. Manhart, I. S. Pandžić, and J. Ahlberg, "Fast facial expression recognition using local binary features and shallow neural networks," *Vis. Comput.*, vol. 36, no. 1, pp. 97–112, Jan. 2020, doi: 10.1007/s00371-018-1585-8.

[2] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 356–370, Jan. 2019, doi: 10.1109/TIP.2018.2868382.

[3] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (Workshops)*, Jun. 2010, p. 2010.

[4] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "Semeval-2018 task 1: Affect in Tweets," in *Proc. 12th Int. Workshop Semantic Eval.*, 2018, pp. 1–17.

[5] S. N. Shivhare and S. Khethawat, "Emotion detection from text," *Comput. Sci., Eng. Appl.*, 2012.

[6] E. Saravia, H.-C.-T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, "CARER: Contextualized affect representations for emotion recognition," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2018.

[7] G. He, X. Liu, F. Fan, and J. You, "Image2audio: Facilitating semi-supervised audio emotion recognition with facial expression image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2020, pp. 912–913.

[8] X. Yu, X. Zhang, Y. Cao, and M. Xia, "VAEGAN: A collaborative filtering framework based on adversarial variational autoencoders," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 4206–4212.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[11] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *Proc. Thematic Workshops ACM Multimedia*, 2017, pp. 349–357.

[12] W. Hao, Z. Zhang, and H. Guan, "CMCGAN: A uniform framework for cross-modal visual-audio mutual generation," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[13] G. Wang, A. Rosenberg, Z. Chen, Y. Zhang, B. Ramabhadran, Y. Wu, and P. Moreno, "Improving speech recognition using consistent predictions on synthesized speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7029–7033.

[14] Y. C. Yoon, S. Y. Park, S. M. Park, and H. Lim, "Image classification and captioning model considering a CAM-based disagreement loss," *ETRI J.*, vol. 42, no. 1, pp. 67–77, Feb. 2020, doi: 10.4218/etrij.2018-0621.

[15] P. P. Liang, R. Salakhutdinov, and L. P. Morency, "Computational modeling of human multimodal language: The mosei dataset and interpretable dynamic fusion," in *Proc. 1st Workshop Grand Challenge Comp. Modeling Hum. Multimodal Lang.*, 2018, pp. 1–23.

[16] W. Dai, Z. Liu, T. Yu, and P. Fung, "Modality-transferable emotion embeddings for low-resource multimodal emotion recognition," 2020, *arXiv:2009.09629*.

[17] J. Liu, S. Chen, L. Wang, Z. Liu, Y. Fu, L. Guo, and J. Dang, "Multimodal emotion recognition with capsule graph convolutional based representation fusion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6339–6343.

[18] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett, "Multiple kernel learning for emotion recognition in the wild," in *Proc. 15th ACM Int. Conf. Multimodal Interact. (ICMI)*, 2013, pp. 517–524.

[19] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3687–3691.

[20] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion," *IEEE Access*, vol. 8, pp. 168865–168878, 2020, doi: 10.1109/ACCESS.2020.3023871.

[21] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 2, 2020, pp. 1359–1367, doi: 10.1609/aaai.v34i02.5492.

[22] Y.-T. Lan, W. Liu, and B.-L. Lu, "Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–9.

[23] H. Zhou, J. Du, Y. Zhang, Q. Wang, Q.-F. Liu, and C.-H. Lee, "Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2617–2629, 2021, doi: 10.1109/TASLP.2021.3096037.

[24] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 6892–6899, doi: 10.1609/aaai.v33i01.33016892.

[25] 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 *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6558–6569, doi: 10.18653/v1/p19-1656.

[26] W. Rahman, M. K. Hasan, S. Lee, A. Bagher Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2359–2369, doi: 10.18653/v1/2020.acl-main.214.

[27] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[28] A. Khare, S. Parthasarathy, and S. Sundaram, "Self-supervised learning with cross-modal transformers for emotion recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2021, pp. 381–388.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[31] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[32] D. Harwath and J. R. Glass, "Learning word-like units from joint audio-visual analysis," 2017, *arXiv:1701.07481*.

[33] J. Qi and Y. Peng, "Cross-modal bidirectional translation via reinforcement learning," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1–7.

[34] B. Yang, B. Shao, L. Wu, and X. Lin, "Multimodal sentiment analysis with unidirectional modality translation," *Neurocomputing*, vol. 467, pp. 130–137, Jan. 2022, doi: 10.1016/j.neucom.2021.09.041.

[35] S. Niu, Y. Jiang, B. Chen, J. Wang, Y. Liu, and H. Song, "Cross-modality transfer learning for image-text information management," *ACM Trans. Manage. Inf. Syst.*, vol. 13, no. 1, pp. 1–14, Mar. 2022, doi: 10.1145/3464324.

[36] D. U. Jo, B. J. Lee, J. Choi, H. Yoo, and J. Y. Choi, "Associative variational auto-encoder with distributed latent spaces and associators," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 11197–11204, doi: 10.1609/aaai.v34i07.6778.

[37] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," 2018, *arXiv:1806.06176*.

[38] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1–11.

[39] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008, doi: 10.1007/s10579-008-9076-6.

[40] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon, "EmotiW 2018: Audio-video, Student engagement and group-level affect prediction," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 653–656.

[41] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391, doi: 10.1371/journal.pone.0196391.

[42] D. Meng, X. Peng, K. Wang, and Y. Qiao, "Frame attention networks for facial expression recognition in videos," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3866–3870.

[43] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition *IEEE/ACM Trans. Audio Speech Langauge Process.*, vol. 28, pp. 2880–2894, 2020, doi: 10.1109/TASLP.2020.3030497.

[44] N. Zeghidour, O. Teboul, F. de Chaumont Quitry, and M. Tagliasacchi, "LEAF: A learnable frontend for audio classification," 2021, *arXiv:2101.08596*.

[45] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[46] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proc. Conf. AAAI Artif. Intell.*, 2018, pp. 5642–5649.

[47] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–9.

[48] M. Xu, F. Zhang, and W. Zhang, "Head fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset," *IEEE Access*, vol. 9, pp. 74539–74549, 2021, doi: 10.1109/ACCESS.2021.3067460.