IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 December 2022, accepted 3 February 2023, date of publication 13 February 2023, date of current version 16 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3244390

RESEARCH ARTICLE

Multi-Label Multimodal Emotion Recognition With Transformer-Based Fusion and Emotion-Level Representation Learning

HOAI-DUY LE, GUEE-SANG LEE, SOO-HYUNG KIM, SEUNGWON KIM, AND HYUNG-JEONG YANG
Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju 61186, South Korea

Corresponding author: Hyung-Jeong Yang (hjyang@jnu.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government (MSIT) under Grant NRF-2020R1A4A1019191; in part by the BK21 Fostering Outstanding Universities for Research (FOUR) funded by the Ministry of Education (MOE), South Korea, and in part by NRF.

ABSTRACT Emotion recognition has been an active research area for a long time. Recently, multimodal emotion recognition from video data has grown in importance with the explosion of video content due to the emergence of short video social media platforms. Effectively incorporating information from multiple modalities in video data to learn robust multimodal representation for improving recognition model performance is still the primary challenge for researchers. In this context, transformer architectures have been widely used and have significantly improved multimodal deep learning and representation learning. Inspired by this, we propose a transformer-based fusion and representation learning method to fuse and enrich multimodal features from raw videos for the task of multi-label video emotion recognition. Specifically, our method takes raw video frames, audio signals, and text subtitles as inputs and passes information from these multiple modalities through a unified transformer architecture for learning a joint multimodal representation. Moreover, we use the label-level representation approach to deal with the multi-label classification task and enhance the model performance. We conduct experiments on two benchmark datasets: Interactive Emotional Dyadic Motion Capture (IEMOCAP) and Carnegie Mellon University Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) to evaluate our proposed method. The experimental results demonstrate that the proposed method outperforms other strong baselines and existing approaches for multi-label video emotion recognition.

INDEX TERMS Multimodal fusion, multi-label video emotion recognition, transformers.

I. INTRODUCTION

Over the past few years, there has been an explosion in short video content from the global rise in short video platforms such as TikTok, YouTube Shorts, and Facebook Reels with the increasing popularity of mobile devices. Online videos have become the predominant data type that users use to share their activities and interact with each other in cyberspace. Inevitably, the scientific and industrial communities have given much attention to video data analysis, especially in video sentiment analysis and emotion recognition [1], [2] because of its applications in diverse fields. People represent their emotions through multiple modalities. Naturally, language, voice, and facial expression are the primary methods by which humans convey their emotions. Moreover, how people combine these methods to express their emotions is very complicated. Under certain circumstances, only relying on information from a single modality to predict human emotions will easily lead to inaccurate predictions.

Let us give a few examples of this phenomenon. Suppose a photographer discovers that his luggage bag containing his camera is missing after landing. He then went to the airport staff to declare and expect to receive the luggage back. After a while of arguing and being given a hard time by the airport

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

staff, he was informed that someone had stolen his baggage and they could not find it. Realizing that he was deliberately wasted his time by airport staff to cover up the fact that they did not ensure the safety of his things. He left with an angry *"Thank you very much"*. His angry emotion will be easily recognized by his facial expressions and tone of voice. On the contrary, if we solely rely on the sentence *"Thank you very much"*, we would obviously think he is grateful. Another example is the case of a woman who has just been promoted and had to work in a foreign office for two years. She is sad because she is about to be separated from her husband and daughter. When asked by her husband about the duration of the trip, she sadly replied, *"It is for two years"*. Analogously, if we use only the text, we will think that the woman's emotion is neutral. By combining her voice and facial expression, we determine that she is unhappy. From these examples, we can be aware of synthesizing information from multiple modalities is crucial to comprehend human emotions fully.

However, designing an effective fusion method for different modalities of video data to obtain a robust joint representation is still a challenging research problem. There have been numerous efforts to design appropriate fusion techniques for incorporating multimodal video information. Early approaches merely concatenated high-level features from all modalities to make a prediction (early fusion) or sum all unimodal decisions with learnable weights (late fusion) to draw the final inference [3], [4]. These fusion methods achieved higher accuracy than unimodal methods, but the improvement was limited because there was no interaction between modalities during training. With the recent advances in deep learning, especially attention mechanisms [5] and transformers [6], later studies were predominantly based on those techniques to explore more sophisticated multimodal fusion methods [7], [8], [9], [10]. These studies commonly process multimodal learning by pairwise or triplet combination input rather than from all the multiple signals concurrently. Nonetheless, it is incompatible with how humans contemporaneously perceive multiple information resources from the video.

In this paper, following the success of the transformers in multimodal deep learning, we present a fusion method built upon the transformer's architecture to effectively fuse multimodal features and learn a joint multimodal representation from the raw video data for multi-label video emotion recognition. Specifically, our proposed model takes raw video frames, audio waveforms, and text transcripts of the video as inputs. The model then extracts high-level features from raw data using appropriate deep neural architectures for each modality and later leverages the multi-head attention mechanism of the transformers to learn a robust joint multimodal representation from the multimodal features. The multi-head attention mechanism in the transformer scans through each element of the input sequence to learn a refined sequence of features that emphasized important elements and faded redundant. We utilize this property in learning the correlation between different modalities from the multimodal input sequence. Furthermore, different from existing transformer-based methods using only the output of the "CLS" token (classification token) to make predictions, we process the entire output sequence of the transformers for classification. We use the trainable query embeddings and the cross-attention in the transformers decoder to learn the emotion-level representations from the joint multimodal features to enhance the multi-label emotion recognition performance.

We evaluate our proposed method on two standard benchmark multimodal emotion recognition datasets: IEMOCAP [11] and CMU-MOSEI [12]. The experiments demonstrate a significant improvement over the strong baseline methods with a gap of +0.2% accuracy and +3.8% F1-score on the IEMOCAP, +2.0% weight accuracy and +0.6% F1-score on the CMU-MOSEI. Overall, the main contributions of our study are as follows:

- We propose a simple but effective multimodal fusion module, which adopts the multi-head attention in the transformer encoder to perform cross-attention and simultaneously integrate informative information at the token level between multimodal features of video data.
- We propose a combination of emotion-level embedding over the fused multimodal features to learn the emotion-related features from multimodal representation. Extensive experiments verify the advantages of this combination in improving performance.
- We conduct extensive experiments and provide thorough ablation studies to demonstrate the effectiveness of our proposed approach to multimodal learning and how our method improves the model performance in video emotion recognition.

The rest of the paper is organized as follows. We provide an overview of prior related studies in Section II. We present the details of our method in Section III. We then describe the extensive experiments and experimental results in Section IV. Finally, we conclude the paper in Section V.

II. RELATED WORK

In this section, we present the related work in two parts: multimodal emotion recognition and multimodal transformers. In each subsection, we first briefly review the progressive existing works and then discuss the ideas for solving remaining constraints.

A. MULTIMODAL EMOTION RECOGNITION

Emotion recognition has been an active research area for many decades. Learning robust representation from multiple modalities has recently become an attractive research direction for boosting emotion recognition performance. Multimodal emotion recognition aims to integrate information from multiple signals such as sound, language, and image in videos for recognizing human emotions.

Most of the previous studies take hand-crafted features extracted by traditional feature extraction algorithms as input to train the deep neural network. Tsai et al. [7] proposed Multimodal Transformer (MulT) which uses pairwise transformers to perform bidirectional cross-attention between visual-textual, visual-audio, and textual-audio for the unaligned multimodal affective recognition task. Dai et al. [13] explored transferring emotion embeddings from textual modality to learn visual and acoustic emotion embeddings and analyzed adaptation in few-shot learning. Hazarika et al. [14] concentrated on enriching multimodal features before the fusion process by introducing MISA which learns modality-specific and modality-invariant representations for multimodal sentiment analysis. In contrast, Han et al. [15] paid attention to constructing a fusion scheme for multimodal data. They considered textual information as the main modality and then designed a Transformer-based fusion network to integrate complementary information from text-visual and text-audio pairs. Besides, Han et al. [16] also presented MultiModal InforMax (MMIM) which applies mutual information concepts in fusing multimodal features for multimodal sentiment analysis. Other approaches received hand-crafted input features for multimodal sentiment analysis and emotion recognition as well [17], [18], [19].

However, it is generally not sensible to train a deep neural network from hand-crafted input features. In this context, Dai et al. [20] recently indicated the limitations of training deep neural networks from hand-crafted input features. They reorganized two benchmark datasets in the multimodal emotion recognition tasks to make training from raw data become feasible. Inspired by their work, we explore a powerful fusion module learned from raw video, audio, and text modalities and further apply it to the task of video emotion recognition.

B. MULTIMODAL TRANSFORMERS

Transformers [6] were originally proposed for the sequence-to-sequence machine translation task and have been widely applied in many other tasks, such as image classification [21], object detection [22], and audio classification [23]. Recently, the transformers have been extended to multimodal deep learning and have been proven effective, especially in learning from textual, visual, and audio modalities of video data. Some early works primarily followed co-attention learning strategies for pairwise modalities. Tan et al. [24] proposed Learning Cross-Modality Encoder Representations from Transformers (LXMERT) framework which is a fully transformer-based network to learn the cross-modality representation of images and languages. They constructed a self-attention learning block for each modality followed by a co-attention learning block to finalize the joint image-text representations. Similarly, Cheng et al. [25] described a co-attention network but for audiovisual synchronization.

Later studies tend to use joint learning with modality-specific transformers trained by contrastive losses for multimodal fusion. Akbari et al. [9] presented a convolution-free transformer-based framework to learn multimodal representations from raw video frames, text transcripts, and audio waveforms of videos in a self-supervised setting. The framework contains separate transformers for each modality and has been trained by a combination of contrastive losses between visual-text and visual-audio pairs. Nagrani et al. [8] introduced a transformers-based approach for fusing audio and visual information in the video by adding bottleneck units to bridge two sequences of visual and audio features before inputting into modality-specific multi-layer transformers. Shvetsova et al. [10] proposed using shared transformer encoders to encode the uni-modal features including text, visual, audio, and pairwise multimodal features comprising text-visual, text-audio, and visual-audio pairs. They then trained the network with combinatorial contrastive losses of pairwise uni-modality and pairs from uni-modal and pairwise modalities. More recently, Mercea et al. [26] modified the standard transformer to perform temporal and enforce cross-attention between visual and audio modalities in the zero-shot learning setting for the video classification task.

In this work, we focus on textual, visual, and audio multimodal fusion for video data. Rather than dividing multiple modalities into pairwise or triplet combinations, we synchronously integrate information from all modalities with a unified transformer-based architecture. We assume that the modalities in a video have inseparable relationships and that humans perceive multiple video modalities simultaneously instead of separately in pairs.

III. PROPOSED METHOD

Given an input video segment V containing multiple modalities including a text transcription, a sequence of video frames, an audio waveform signal, and a pre-defined set of K emotions, multi-label emotion recognition is to predict whether each emotion is present in the video. In this section, we provide a detailed description of our proposed approach to solve the above problem.

Figure 1 illustrates the overall architecture of our proposed method for multi-label multimodal emotion recognition taking a video clip as an input and outputting the revealed emotions in the video. The model consists of three modules: the feature extraction module, the multimodal fusion module, and the emotion-level embedding module. First, we employ a feature extraction module to encode the text transcription, video frames, and audio signal inputs into hidden representations. Then, in the fusion module, we leverage the multi-head attention in transformer encoders to fuse the features of multiple input modalities. After that, we construct the emotion-level embedding module using transformer decoders to learn the emotion-level representations from the fused multimodal representation. Finally, we apply a fully feed-forward layer followed by a sigmoid activation layer to make predictions and apply weighted binary cross-entropy loss to train the network. The detail of the proposed method is described in the below subsections.

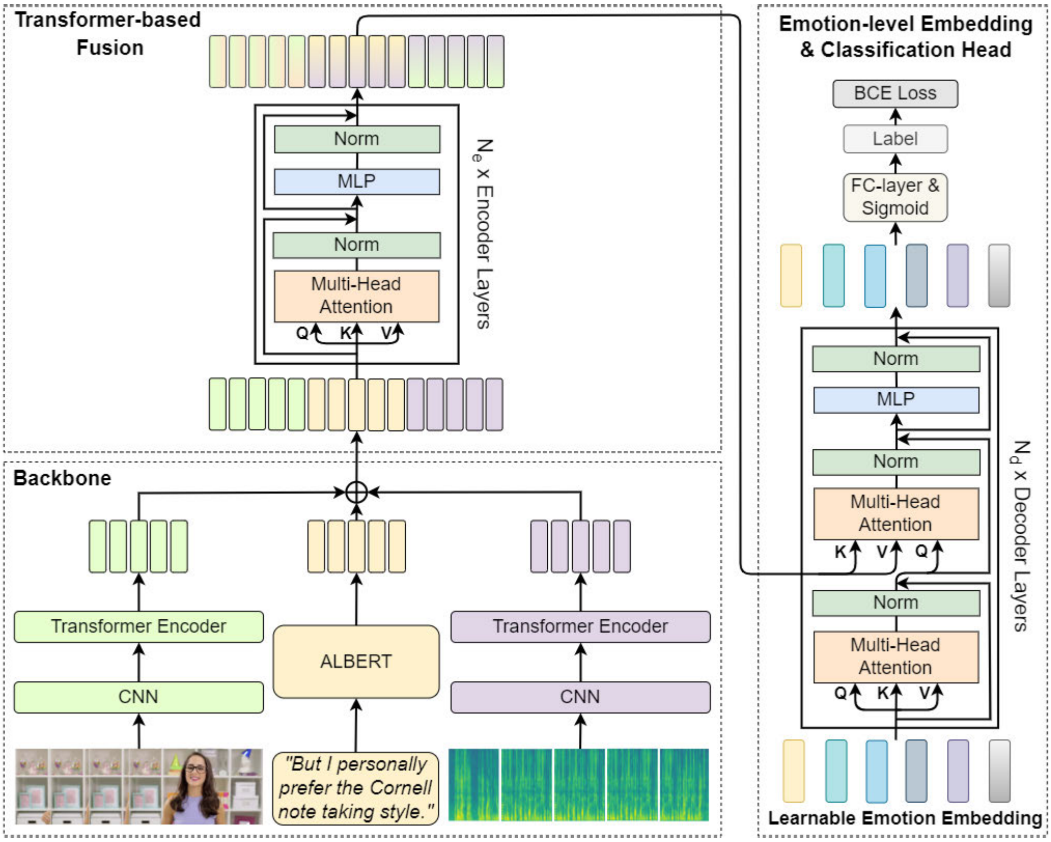

FIGURE 1. Architecture of our proposed method. It consists of three main modules: (1) a backbone module containing three feature extractors for textual, visual, and acoustic modalities, (2) a Transformer-based fusion module that attends and fuses multimodal information, and (3) an emotion-level embedding and classification head module that matches the fused multimodal features to emotion-level representations and outputs the final emotion predictions.

A. FEATURE EXTRACTION

This study focuses on designing multimodal fusion and enriching representation learning for video data to improve model performance. Thus, we use the identical feature extraction following the baseline [20] rather than proposing a new extractor. Dai et al. [20] proposed a multimodal model learned from raw video data for multimodal emotion recognition. First, a feature extractor is employed to extract features from each modality. Then the late fusion approach is applied to fuse multimodal information and predict the final emotional classes. In detail, the feature extraction module leveraged in [20] and our work consists of three modality-specific extractors. We use a pre-trained BERT-based model to extract a set of word embeddings from textual modality. For visual (video frames) and acoustic (Mel spectrogram chunks) modalities, we utilize two individual CNN networks (trained from scratch) as the backbones of each modality. We enable fusing multimodal features by including a projection network containing multiple fully-connected layers followed by a non-linear activation function for each modality to map the multimodal features to the same size. Furthermore, we use transformer encoders to capture the temporal information of the sequence of hidden features from video frames and audio spectrograms. Consequently, we obtain three sequences of hidden representations from the textual, visual, and acoustic modalities denoted as $T \in \mathbb{R}^{n_t \times d}$, $I \in \mathbb{R}^{n_i \times d}$, and $A \in \mathbb{R}^{n_a \times d}$, respectively. In which, n_t, n_i, and n_a are the numbers of words in the transcription, the number of sampled video frames, and the number of Mel spectrogram chunks, respectively; d is the size of feature dimensionality. For the textual modality, $T \in \mathbb{R}^{n_t \times d}$ consists of n_t word embeddings extracted from the pre-trained BERT-based model and projected to d size. For the visual and acoustic modalities, $I \in \mathbb{R}^{n_i \times d}$ and $A \in \mathbb{R}^{n_a \times d}$ contain n_i and n_a feature vectors of size d captured from videos frames and audio spectrogram chunks, respectively.

B. TRANSFORMER-BASED MULTIMODAL FUSION MODULE

Unlike previous transformer-based multimodal fusion methods, we propose synchronously fusing features from different modalities rather than dividing combinations of possible pairwise modalities. We first summarize the multi-head attention in the transformer [6] and then describe our proposed extension to multimodal fusion for video data. Given an input sequence $S \in \mathbb{R}^{n \times d}$ containing n vectors of size d, the multi-head self-attention block in the transformer parallelly

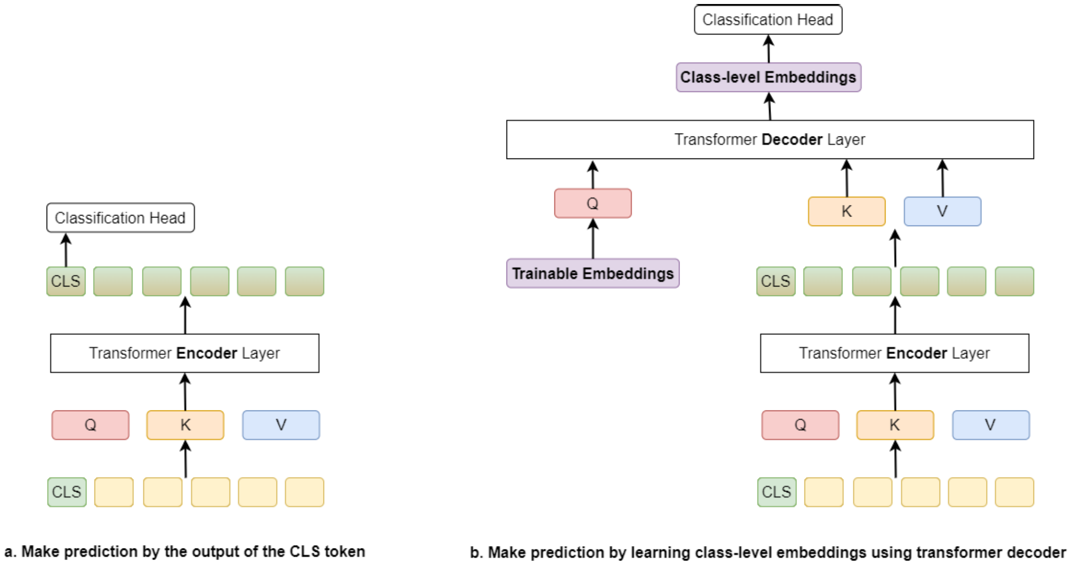
Classification Head
CLS
Transformer Encoder Layer
Q
K
V
CLS
a. Make prediction by the output of the CLS token
Classification Head
Class-level Embeddings
Transformer Decoder Layer
Q
Trainable Embeddings
K
V
CLS
Transformer Encoder Layer
Q
K
V
CLS
b. Make prediction by learning class-level embeddings using transformer decoder

FIGURE 2. Illustration of learning class-level embeddings (Figure 2. b) compared to using the output of the CLS token (Figure 2. a) for classification.

projects S to multiple sets of three components named query $Q_i \in \mathbb{R}^{n \times d_k}$, key $K_i \in \mathbb{R}^{n \times d_k}$, and value $V_i \in \mathbb{R}^{n \times d_v}$ in h different subspaces (h is the number of heads, $d_k = d/h$, and usually $d_k = d_v$). On each of these sets of projected queries, keys, and values, a single attention function is performed as follows:

$$Attention(Q_i, K_i, V_i) = softmax(\frac{Q_i K_i^T}{\sqrt{d_k}})V_i.$$
$$softmax(x_i) = \frac{exp(x_i)}{\Sigma_j exp(x_j)} \tag{1}$$

The dot product QK^T is the form of the similarity measure and the $Attention(Q, K, V)$ is the sum weighted by attention weight (softmax score). The final joint representation is obtained by averaging all attention head outputs with trainable weights:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^o,$$
$$head_i = Attention(Q_i, K_i, V_i), \ i \in 1, \ldots, h, \tag{2}$$

where $W^o \in \mathbb{R}^{hd_v \times d}$ represents the learnable parameters. A single dot product attention enables the model to scan through each element in the input sequence and learns which elements it should attend to. The multi-head attention enables this process to be executed from different representation subspaces. In other words, the transformers provide a mechanism to selectively accumulate information from the entire input sequence with regard to the output. Furthermore, the multi-head attention is enduring with the order of vectors in the input sequence. Therefore, it is naturally suitable to fuse multimodal information by applying multi-head attention over the input sequence which is the order-agnostic combination of features from multiple modalities.

After passing the feature extraction module, we obtain three sequences of hidden representations $T \in \mathbb{R}^{n_t \times d}$, $I \in \mathbb{R}^{n_i \times d}$, and $A \in \mathbb{R}^{n_a \times d}$ from multiple modalities including text transcription, image frames, and acoustic signals, respectively. We then concatenate them into a unified sequence of multimodal features added with a classification token ([CLS] token) at the start and use it as the input sequence for the fusion module. We adopt the vanilla transformer [6] encoder and stack at L_e multiple blocks to construct the multimodal fusion module. The standard transformer encoder consists of a multi-head self-attention layer (MSA), normalization layers (Norm), and a position-wise feed-forward network (FFN). The fused multimodal representation $F^i \in \mathbb{R}^{n_s \times d}$ ($n_s = n_t + n_i + n_a$) at block i is calculated as follows:

$$F^i_{temp} = Norm(MSA(F^{i-1}) + F^{i-1}), \tag{3}$$
$$F^i = Norm(FFN(F^i_{temp}) + F^i_{temp}), \tag{4}$$

in which $i \in 1, \ldots, L_e$ and $F^0 = concat(E_{cls}, T, I, A)$.

C. EMOTION-LEVEL EMBEDDING MODULE

In contrast to previous transformer-based works which commonly use the output of the classification token ("[CLS]" token) to perform classification with linear layers, we make use of the entire outputted sequence from transformer encoders of the fusion module to enrich features for the multi-label emotion recognition task as illustrated in Figure 2. Rather than learning a unique representation and then using it to make predictions for all emotions, we adopt the idea of learning multiple embeddings, in which each embedding is oriented toward each specific emotion. We leverage the cross-attention in the transformer decoder to pool multiple emotion-level embeddings for a single video inspired by [22], [27], and [28].

The emotion-level embedding module takes the sequence of features from the output of the fusion module $F \in \mathbb{R}^{n_s \times d}$ as input and generates the emotion-level representation $E \in \mathbb{R}^{C \times d}$ (C equal to the number of emotional classes and d is the feature dimensional size) for the video. First, a set of emotion-specific embeddings $E_0 \in \mathbb{R}^{C \times d}$ is randomly initialized and used to project the query vector Q. It will be learned during the training process. Simultaneously, the sequence of refined multimodal features outputted from the fusion module is used to project K and V vectors. The video emotion-level representation is then learned using a series of N_d transformer decoders:

$$E_i^{\dagger} = Norm(E_{i-1} + MHA(E_{i-1}, E_{i-1}, E_{i-1})) \tag{5}$$

$$E_i^{\dagger\dagger} = Norm(E_i^{\dagger} + MHA(E_i^{\dagger}, F, F)) \tag{6}$$

$$E_i = Norm(E_i^{\dagger\dagger} + FFN(E_i^{\dagger\dagger})) \tag{7}$$

where *Norm*, *MHA*, and *FFN* are the normalization layer, Multi-Head Attention layer, and feed-forward network, respectively; $i \in \{1, \ldots, N_d\}$.

D. OBJECTIVE FUNCTION

Given an input video labeled by a multi-hot vector $Y = [y_0, y_1, \ldots, y_C]$, $y_i \in \{0, 1\}$, our proposed model outputs the confidence scores of all classes $Y = [p_0, p_1, \ldots, p_C]$, $p_i \in [0, 1]$. A confidence score is a probability given by the model to represent how confident the model assigns the input to a certain class. The network is trained as end-to-end learning with binary cross-entropy loss (BCE) to adapt to the multi-label classification. Because of the class imbalance problem in the datasets, we add weights to the loss of positive samples. The loss for each training sample is formulated as follows:

$$L(y, p) = \sum -\frac{1}{C}[w_i.y_i.log(p_i) + (1 - y_i).log(1 - p_i)], \tag{8}$$

$$w_i = \frac{n_i}{p_i}, \tag{9}$$

where n_i and p_i denote the number of negative and positive samples of class i, respectively. For mini-batch learning, the total loss is the average of all sample losses in the batch.

IV. EXPERIMENTS

A. DATASETS AND METRICS

We conduct experiments on two benchmarked datasets including Interactive Emotional Dyadic Motion Capture (IEMOCAP) [11] and CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [12]. We re-use the reorganized version of these datasets from Dai et al. work [20] instead of the original version because the input of our method is raw video. We also follow the split for training, validation, and testing in [20]. Table 1 and table 2 show the statistics of both IEMCAP and CMU-MOSEI datasets.

TABLE 1. The detailed statistics in IEMOCAP and CMU-MOSEI datasets. *"t, v, a"* stands for textual, visual, and audio modalities, respectively.

Dataset	IEMOCAP	CMU-MOSEI
Modality	t,v,a	t,v,a
Training size	5162	14524
Validation size	737	1765
Testing size	1481	4188
Total samples	7380	20477
Emotion labels	anger, excited, frustrated, happiness, neutral, sadness	anger,disgust,fear, happiness, sadness, surprise

TABLE 2. The detailed emotional label distribution of IEMOCAP and CMU-MOSEI datasets.

Label	IEMOCAP	Label	CMU-MOSEI
Angry	1103	Angry	4600
Excited	1041	Disgusted	3755
Frustrated	1849	Fear	1803
Happy	595	Happy	10752
Neutral	1708	Sad	5601
Sad	1084	Surprised	2055

1) IEMOCAP [11]

contains 151 recorded dialogue videos. In each video, two speakers have a dyadic conversation with multiple utterances. Originally, the dataset is annotated by nine emotional labels. Due to the imbalance problem, [20] preserves only six categories among them including *angry, happy, excited, sad, frustrated*, and *neutral*. The dialogue videos are sliced at the utterance level into 7380 sub-clips. Because of the shortage of identifiers for data samples in the original training-validation-testing split from [11] and [20] constructed a new split from sliced videos with a 70%-10%-20% ratio for training, validation, and testing configuration, respectively.

2) CMU-MOSEI [12]

is the largest dataset for multimodal sentiment analysis and emotion recognition tasks. The dataset consists of 23,259 utterance-video segments sampled from 3,837 YouTube videos from 1,000 distinct speakers. It is labeled into six emotion categories: *happy, sad, angry, fearful, disgusted*, and *surprised*. After cleaning misaligned and mismatched data samples, 20,477 videos remain in the dataset. The new dataset split is done by following the CMU-MOSEI split for the sentiment analysis task.

3) EVALUATION METRICS

We use the standard accuracy and F1-score as evaluation metrics for the IEMOCAP dataset consistent with previous works. For CMU-MOSEI, because of the imbalance of positive and negative samples in each emotion class, weighted accuracy is used instead of standard accuracy; besides, F1-score is also calculated to evaluate the model performance. The F1-score and the weighted accuracy are defined as follows:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN}, \tag{10}$$

TABLE 3. Results on the IEMOCAP dataset. Comparison with strong baselines and existing methods. ∗: excerpted from previous papers, †: reproduced from open-source code with hyper-parameter provided in the original paper.

Method	Angry		Excited		Frustrated		Happy		Neutral		Sad		Avg.	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Audio - VGG19	82.1	46.2	55.0	30.3	45.4	42.4	12.8	17.2	48.6	41.4	77.9	48.8	53.6	37.7
Image - VGG19	80.9	53.8	84.3	56.3	68.3	53.9	90.1	43.7	75.1	55.3	86.6	55.8	80.9	53.1
Text - ALBERT	86.0	57.9	86.8	56.8	70.1	54.7	90.1	40.5	73.5	49.0	87.0	58.0	82.2	52.8
Late Fusion - LSTM ∗	71.2	49.4	79.3	57.2	68.2	51.5	67.2	37.6	66.5	47.0	78.2	54.0	71.8	49.5
Late Fusion - Trans ∗	81.9	50.7	85.3	57.3	60.5	49.3	85.2	37.6	72.4	49.7	87.4	57.4	78.8	50.3
EmoEmbs [13] ∗	65.9	48.9	73.5	58.3	68.5	52.0	69.6	38.3	73.6	48.7	80.8	53.0	72.0	49.8
MulT [7] ∗	77.9	60.7	76.9	58.0	72.4	57.0	80.0	46.8	74.9	53.7	83.5	65.4	77.6	56.9
FE2E [20] †	89.4	62.5	86.6	61.0	75.7	**59.3**	**91.7**	34.9	**79.2**	56.6	**91.3**	68.5	85.7	57.1
Ours	**90.1**	**66.8**	**88.5**	**66.8**	**77.7**	57.0	90.5	**48.5**	78.1	**56.6**	90.7	**69.6**	**85.9**	**60.9**

$$WAcc = \frac{TP \times N/P + TN}{2N}, \tag{11}$$

where TP (resp. TN) stands for true positive (resp. true negative), FP (resp. FN) denotes false positive (resp. false negative), and N (resp. P) is total negative (resp. positive) samples.

B. IMPLEMENTATION DETAILS

1) MODEL SETTING

We implement the proposed model with Pytorch [29] framework v.1.8.1. To ensure comparability, we follow [20] to build up the backbone for the feature extraction module. In particular, we use a pre-trained ALBERT-base model [30] to extract word embeddings from text transcription. The max length of the text is limited to 50. For all experiments, the video frames are sampled every 1s and then passed through a pre-trained MTCNN [31] model to detect and crop face regions. The cropped faces are resized to 64 × 64 and used as input for the visual backbone. For the acoustic modality, the audio signals are converted to the Mel spectrogram form and sliced into a sequence of chunks with a time window of 400 ms. We construct the VGG19 [32] architecture trained from scratch as the backbone for video frames and audio Mel spectrogram chunks. The projection networks have two fully-connected layers with ReLU activation. The number of encoders and decoders for the image encoder, audio encoder, fusion module, and emotion-level embedding network used for each dataset are clarified in the ablation study. All used transformer encoders and decoders have 4 heads ($h = 4$) with a hidden size of 256 ($d = 256$) and a feed-forward network dimensional of 2048.

2) MODEL TRAINING

We trained the network using the Adam [33] optimizer with a cosine decay learning rate schedule [34]. The initial pre-trained ALBERT model's learning rate is 5×10^{-5}, and the initial learning rate for the other layer's weights is 5×10^{-4}. We set the batch size to 32 and train the model on 1 RTX8000 GPU with 48GB RAM. We train 10 epochs for the CMU-MOSEI and 25 epochs for the IEMOCAP dataset.

3) BASELINES

We compare our method with the following baselines and existing approaches. First, we compare with strong unimodal methods trained from raw data including ALBERT for textual, and VGG19 for visual and audio. Second, we compare with multimodal methods trained from hand-crafted features including the Emotion Embeddings (EmoEmbs) model [13], and the Multimodal Transformer (MulT) model [7]. They are considered strong baselines. Finally, we compare our method with the FE2E [20] which merely adopts late fusion for fusing multimodal features extracted from raw video data. We consider the FE2E as our main competitor because it is the first work using end-to-end training manner from raw data for video emotion recognition.

C. MAIN RESULTS

Table 3 provides the quantitative results of our proposed method compared with other strong baselines on the IEMOCAP dataset. Our approach performs better than the baselines and previous methods, with an F1 score of 60.9% and an accuracy of 85.9%. Table 4 summarizes the comparative results of our model and other competitive existing methods on the CMU-MOSEI dataset. Our network outperforms all other approaches in terms of the F1 score and achieves a conspicuous accuracy improvement with a gap of +2% compared to the FE2E. From both tables, we can further observe that our proposed approach significantly enhances performance on minority emotions (anger, excitement, and happiness on the IEMOCAP; disgust, fear, and surprise on the CMU-MOSEI). These results demonstrate the effectiveness of our proposed network in fusing multimodal video data and learning powerful representations for multi-label emotion recognition. Moreover, the results reinforce that multimodal learning is superior to unimodal and that training from raw data is better than hand-crafted features.

D. ABLATION STUDY

We carried out ablation studies to evaluate the influence of the transformer-based fusion module and emotion-level embedding module on the model performance. To examine the impacts of a module, we conduct the same experimental

TABLE 4. Results on the CMU-MOSEI dataset. Comparison with strong baselines and existing methods. ∗: excerpted from previous papers, †: reproduced from open-source code with hyper-parameter provided in the original paper.

Method	**Angry**		**Disgusted**		**Fear**		**Happy**		**Sad**		**Surprised**		**Avg.**	
	WAcc	F1	WAcc	F1	WAcc	F1	WAcc	F1	WAcc	F1	WAcc	F1	WAcc	F1
Audio - VGG19	53.9	40.5	61.0	35.7	59.0	19.6	50.0	69.3	61.2	45.8	58.4	21.7	57.2	38.8
Image - VGG19	58.9	38.2	63.2	37.6	59.1	21.8	55.7	70.3	56.2	42.8	53.0	17.9	57.7	38.1
Text - ALBERT	65.9	48.4	74.0	56.0	62.8	27.0	62.3	72.0	60.2	45.3	60.9	26.0	64.3	45.8
Late Fusion - LSTM ∗	64.5	47.1	70.5	49.8	61.7	22.2	61.3	73.2	63.4	47.2	57.1	20.6	63.1	43.3
Late Fusion - Trans ∗	65.3	47.7	74.4	51.9	62.1	24.0	60.6	72.9	60.1	45.5	62.1	24.2	64.1	44.4
EmoEmbs [13] ∗	66.8	49.4	69.6	48.7	63.8	23.4	61.2	71.9	60.5	47.5	63.3	24.0	64.2	44.2
MulT [7] ∗	64.9	47.5	71.6	49.3	62.9	25.3	**67.2**	**75.4**	64.0	48.3	61.4	25.6	65.4	45.2
FE2E [20] †	66.9	49.5	75.4	**57.2**	63.8	27.1	61.9	72.3	**65.6**	**49.3**	61.5	26.9	65.8	47.0
Ours	**67.5**	**50.2**	**76.3**	57.0	**69.0**	**29.0**	63.0	72.6	65.5	49.2	**65.7**	**27.6**	**67.8**	**47.6**

TABLE 5. Component-wise ablation analysis on both IEMOCAP and CMU-MOSEI datasets.

Model	**IEMOCAP**		**CMU-MOSEI**	
	Acc	F1	WAcc	F1
Text - ALBERT	82.23	52.80	64.34	45.77
Image - VGG19	80.86	53.12	57.68	38.08
Spectrogram - VGG19	53.63	37.74	57.25	38.77
Early Fusion	84.72	57.19	65.49	45.49
+ Transformer-based Fusion	**85.02** (+0.3)	**58.58** (+1.39)	**66.35** (+0.86)	**46.39** (+0.9)
+ Transformer-based Fusion + Emotion-level Embedding	**85.92** (+0.9)	**60.89** (+2.31)	**67.81** (+1.46)	**47.60** (+1.21)

TABLE 6. Ablation analysis on the effects of the number of transformer encoders and decoders on both IEMOCAP and CMU-MOSEI datasets. "ImgEnc", "SpecEnd", and "FusionEnc" stand for the number of transformer encoders in image feature extraction, audio feature extraction, and fusion module, respectively. "EmoDec" implies the number of transformer decoders in the emotion-level embedding network.

Img Enc	Spec Enc	Fusion Enc	Emo Dec	IEMOCAP		CMU-MOSEI	
				Acc	F1	WAcc	F1
1	1	1	1	83.79	58.58	66.95	47.10
1	1	2	2	84.82	61.88	66.23	46.92
2	2	2	2	**85.92**	**60.89**	67.11	47.27
2	2	4	4	84.23	59.37	66.92	47.53
4	4	4	4	83.72	57.90	**67.81**	**47.60**
4	4	6	6	84.04	58.15	66.70	47.86

settings using the proposed method with and without the component. The contributions of these modules are depicted in Table 5. Particularly, in the first test case, we exclude the emotion-level embedding module to explore the effects of the fusion module. By using the fusion module, we improve the Acc and F1 score by +0.3% and +1.39% on the IEMOCAP and the Wacc and F1 score by +0.86% and +0.9% on the CMU-MOSEI. For the second test case, we attach the emotion-level embedding module to the fusion module to train the model. As can be observed in Table 5, the model performance is improved more. On the IEMOCAP, adding emotion-level embedding achieves 0.9% Acc and 2.31% F1 score improvement. On the CMU-MOSEI, the Wacc and F1 score increases are 1.46% and 1.21%, respectively. In summary, the proposed network combined transformer-based fusion module and emotion-label embedding module achieves the highest performance in terms of both accuracy and F1 score on both IEMOCAP and CMU-MOSEI datasets.

We further provide ablation studies in order to select the optimal number of transformer encoder layers for the feature extraction module and fusion module, and the optimal number of transformer decoder layers for the emotion-level embedding network. Table 6 shows the experimental results for both the IEMOCAP and CMU-MOSEI datasets. Our method achieves the highest performance with 2 layers for transformer encoders and decoders in all modules on the IEMOCAP and 4 layers for all modules on the CMU-MOSEI. These configurations are quantitatively consistent with the size of the datasets. The larger the dataset, the more layers are needed.

V. CONCLUSION

In this paper, we presented an end-to-end Transformer-based fusion and emotion-level representation learning method for multi-label multimodal emotion recognition. Our approach takes a raw video as input rather than hand-crafted features used in most other previous works. We leverage the multi-head attention in the transformers to concurrently fuse the multimodal features of video data at multiple layers. Moreover, we proposed learning emotion-level representations from fused multimodal features to improve model performance in the multi-label recognition task. We demonstrated the effectiveness of our proposed model on two standard benchmark datasets. The experimental results show that our model outperforms other strong baselines and previous existing methods. The provided ablation study clearly illustrates the contribution of the fusion module and the emotion-level embeddings module to the model performance improvement.

During the experimental process, we observe that visual modality (video frames) is the most computational expense. This makes the training process more time-consuming and affects usability in real-world applications. In fact, in a sequence of video frames, there are a lot of redundant and almost identical frames. In future work, we will investigate learning to filter irrelevant and duplicated frames to reduce the computational cost so that the model can be practical in real-world scenarios.

REFERENCES

[1] Z. Li, Y. Fan, B. Jiang, T. Lei, and W. Liu, ''A survey on sentiment analysis and opinion mining for social multimedia,'' *Multimedia Tools Appl.*, vol. 78, no. 6, pp. 6939–6967, 2019.

[2] R. Kaur and S. Kautish, ''Multimodal sentiment analysis: A survey and comparison,'' in *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines*, 2022, pp. 1846–1870.

[3] S. Chen, X. Li, Q. Jin, S. Zhang, and Y. Qin, ''Video emotion recognition in the wild based on fusion of multimodal features,'' in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 494–500.

[4] J. D. S. Ortega, P. Cardinal, and A. L. Koerich, ''Emotion recognition using fusion of audio and video features,'' in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Oct. 2019, pp. 3847–3852.

[5] M.-T. Luong, H. Pham, and C. D. Manning, ''Effective approaches to attention-based neural machine translation,'' 2015, *arXiv:1508.04025*.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, ''Attention is all you need,'' in *Advances in Neural Information Processing Systems*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[7] 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, p. 6558.

[8] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, ''Attention bottlenecks for multimodal fusion,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 14200–14213.

[9] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, ''VATT: Transformers for multimodal self-supervised learning from raw video, audio and text,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24206–24221.

[10] N. Shvetsova, B. Chen, A. Rouditchenko, S. Thomas, B. Kingsbury, R. Feris, D. Harwath, J. Glass, and H. Kuehne, ''Everything at once—Multi-modal fusion transformer for video retrieval,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20020–20029.

[11] 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.

[12] A. Bagher 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*, vol. 1. Melbourne, QC, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2236–2246. [Online]. Available: https://aclanthology.org/P18-1208

[13] W. Dai, Z. Liu, T. Yu, and P. Fung, ''Modality-transferable emotion embeddings for low-resource multimodal emotion recognition,'' 2020, *arXiv:2009.09629*.

[14] D. Hazarika, R. Zimmermann, and S. Poria, ''MISA: Modality-invariant and -Specific representations for multimodal sentiment analysis,'' in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1122–1131.

[15] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.-P. Morency, and S. Poria, ''Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis,'' in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2021, pp. 6–15.

[16] W. Han, H. Chen, and S. Poria, ''Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.* Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 9180–9192. [Online]. Available: https://aclanthology.org/2021.emnlp-main.723

[17] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, ''Tensor fusion network for multimodal sentiment analysis,'' 2017, *arXiv:1707.07250*.

[18] 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,'' 2018, *arXiv:1806.00064*.

[19] A. Shenoy and A. Sardana, ''Multilogue-Net: A context aware RNN for multi-modal emotion detection and sentiment analysis in conversation,'' 2020, *arXiv:2002.08267*.

[20] W. Dai, S. Cahyawijaya, Z. Liu, and P. Fung, ''Multimodal end-to-end sparse model for emotion recognition,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.* Toronto, ON, Canada: Association for Computational Linguistics, 2021, pp. 5305–5316. [Online]. Available: https://aclanthology.org/2021.naacl-main.417

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, ''An image is worth 16×16 words: Transformers for image recognition at scale,'' 2020, *arXiv:2010.11929*.

[22] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, ''End-to-end object detection with transformers,'' in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 213–229.

[23] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, ''SSAST: Self-supervised audio spectrogram transformer,'' in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 10, 2022, pp. 10699–10709.

[24] H. Tan and M. Bansal, ''LXMERT: Learning cross-modality encoder representations from transformers,'' 2019, *arXiv:1908.07490*.

[25] Y. Cheng, R. Wang, Z. Pan, R. Feng, and Y. Zhang, ''Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning,'' in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3884–3892.

[26] O.-B. Mercea, T. Hummel, A. Koepke, and Z. Akata, ''Temporal and cross-modal attention for audio-visual zero-shot learning,'' in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 488–505.

[27] R. You, Z. Guo, L. Cui, X. Long, Y. Bao, and S. Wen, ''Cross-modality attention with semantic graph embedding for multi-label classification,'' in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 12709–12716.

[28] S. Liu, L. Zhang, X. Yang, H. Su, and J. Zhu, ''Query2Label: A simple transformer way to multi-label classification,'' 2021, *arXiv:2107.10834*.

[29] A. Paszke, S. Gross, F. Massa, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, and A. Köpf, ''PyTorch: An imperative style, high-performance deep learning library,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[30] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, ''ALBERT: A lite BERT for self-supervised learning of language representations,'' in *Proc. 8th Int. Conf. Learn. Represent. (ICLR)*, Addis Ababa, Ethiopia, Apr. 2020, pp. 1–17. [Online]. Available: https://openreview.net/forum?id=H1eA7AEtvS

[31] J. Xiang and G. Zhu, ''Joint face detection and facial expression recognition with MTCNN,'' in *Proc. 4th Int. Conf. Inf. Sci. Control Eng. (ICISCE)*, 2017, pp. 424–427.

[32] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' 2014, *arXiv:1409.1556*.

[33] D. P. Kingma and J. Ba, ''Adam: A method for stochastic optimization,'' 2014, *arXiv:1412.6980*.

[34] I. Loshchilov and F. Hutter, ''SGDR: Stochastic gradient descent with warm restarts,'' 2016, *arXiv:1608.03983*.

HOAI-DUY LE received the B.S. degree in electronics and telecommunication engineering from the Ho Chi Minh University of Technology, Vietnam, in 2018. He is currently pursuing the master's degree with the Department of Artificial Intelligence Convergence, Chonnam National University, South Korea. His research interests include multimodal deep learning, video understanding, affective computing, data mining, and social media analytics.

GUEE-SANG LEE received the B.S. degree in electrical engineering and the M.S. degree in computer engineering from Seoul National University, South Korea, in 1980 and 1982, respectively, and the Ph.D. degree in computer science from Pennsylvania State University, in 1991. He is currently a Professor with the Department of Artificial Intelligence Convergence, Chonnam National University, South Korea. His research interests include image processing, computer vision, and video technology.

SOO-HYUNG KIM received the B.S. degree in computer engineering from Seoul National University, in 1986, and the M.S. and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology, in 1988 and 1993, respectively. Since 1997, he has been a Professor with the Department of Artificial Intelligence Convergence, Chonnam National University, South Korea. His research interests include pattern recognition, document image processing, medical image processing, and ubiquitous computing.

SEUNGWON KIM received the bachelor's and master's degrees from the University of Tasmania, in 2008 and 2010, respectively, and the Ph.D. degree from the HIT Lab NZ, New Zealand, in 2016, under the supervision of Prof. M. Billinghurst. During his Ph.D. study, he developed a remote collaboration system supporting stabilized sketch and pointer cues. He is currently a Professor with the Department of Artificial Intelligence Convergence, Chonnam National University, South Korea. His research interests include remote collaboration using augmented virtual communication cues and sharing experiences between distance users.

HYUNG-JEONG YANG received the B.S., M.S., and Ph.D. degrees from Chonbuk National University, South Korea. She is currently a Professor with the Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, South Korea. Her main research interests include multimedia data mining, medical data analysis, social network service data mining, and video data understanding.