Received March 21, 2022, accepted May 4, 2022, date of publication May 11, 2022, date of current version May 18, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3174215

VAE-Based Adversarial Multimodal Domain Transfer for Video-Level Sentiment Analysis

YANAN WANG[1,2], (Student Member, IEEE), JIANMING WU[1], KAZUAKI FURUMAI[1], SHINYA WADA[1], AND SATOSHI KURIHARA[2], (Member, IEEE)

[1]Data Intelligence Division, KDDI Research Inc., Saitama 356-8502, Japan
[2]School of Science for Open and Environmental Systems, Keio University, Tokyo 223-8522, Japan

Corresponding author: Yanan Wang (wa-yanan@kddi.com)

ABSTRACT Video-level sentiment analysis is a challenging task and requires systems to obtain discriminative multimodal representations that can capture difference in sentiments across various modalities. However, due to diverse distributions of various modalities and the unified multimodal labels are not always adaptable to unimodal learning, the distance difference between unimodal representations increases, and prevents systems from learning discriminative multimodal representations. In this paper, to obtain more discriminative multimodal representations that can further improve systems' performance, we propose a VAE-based adversarial multimodal domain transfer (*VAE-AMDT*) and jointly train it with a multi-attention module to reduce the distance difference between unimodal representations. We first perform variational autoencoder (VAE) to make visual, linguistic and acoustic representations follow a common distribution, and then introduce adversarial training to transfer all unimodal representations to a joint embedding space. As a result, we fuse various modalities on this joint embedding space via the multi-attention module, which consists of self-attention, cross-attention and triple-attention for highlighting important sentimental representations over time and modality. Our method improves F1-score of the state-of-the-art by **3.6%** on MOSI and **2.9%** on MOSEI datasets, and prove its efficacy in obtaining discriminative multimodal representations for video-level sentiment analysis.

INDEX TERMS Multimodal representation learning, domain adaptation, variational auto-encoder (VAE), adversarial training.

I. INTRODUCTION

Video-level sentiment analysis is a task to predict people's sentiment intensity with a given video clip. It is an essential task for achieving high-level artificial intelligence (AI), and is expected to be applied to dialogue agents, virtual reality and social robotics, and so on [1]. To let AI systems have a better understanding of people's sentiment, existing methods fuse multimodal representations obtained from video frame (image), text and audio, and predict sentiment intensity by doing regression analysis [7], [9]. How to obtain discriminative multimodal representations that can capture difference in sentiments across various modalities is a core issue for video-level sentiment analysis [2], [10], [11]. However, due to diverse distributions of various modalities (*e.g.*, one same sentiment intensity corresponds to different unimodal representations.) and the unified multimodal labels are not always adaptable to unimodal learning (*e.g.*, an unified multimodal label is *highly negative*, but text represents *neutral*), the distance difference between unimodal representations increases, and prevents systems from learning discriminative multimodal representations. Mai *et al.* [16] propose adversarial encoder-decoder-classifier framework to reduce modality gap by using adversarial training [3], [30], and Yu *et al.* [15] design an unimodal label auto generation module to better learn unimodal representations for multimodal fusion. These two methods reduce the distance difference between unimodal representations via different approach, aim to map various modalities in a joint embedding space so that the model can easily learn a common classifier. However, from the evaluation result, their efficacy is limited on the small and imbalanced sentiment dataset.

In this paper, to obtain more discriminative multimodal representations that can further improve the performance of video-level sentiment analysis, as shown in Fig. 1, we propose a VAE-based adversarial multimodal domain transfer (*VAE-AMDT*) to better reduce the distance difference between unimodal representations and transfer various

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

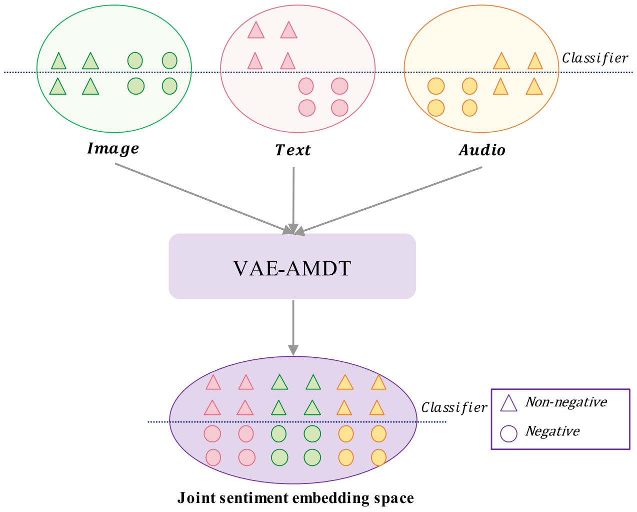


FIGURE 1. A conceptual diagram illustrates distribution of various modalities is diversity. ***VAE-AMDT*** is designed to transfer unimodal representations to a joint sentiment embedding space. As a result, we obtain discriminative sentiment multimodal representations and make it easier to predict sentiment intensity. "△" and "○" indicate "non-negative" and "negative" respectively.

modalities to a joint embedding space, so that the model can easily learn discriminative multimodal representations and find an effective classifier over various modalities. Variational auto-encoder (VAE) is an auto-encoder whose training is regularised so that the distributions returned by its encoder are enforced to be close to a standard normal distribution [4], [5]. We perform it with visual, linguistic and acoustic modality respectively to make encoded latent representations follow a common distribution so that the modality gap can be reduced. Furthermore, motivated by [16], we introduce discriminator trained with adversarial loss to classify encoded latent representation of target modality as true but others as false. As a result, we can better transfer encoded latent representations from various modalities to a joint embedding space as shown in Fig. 1. And then, we jointly train *VAE-AMDT* with a multi-attention module on this joint embedding space to learn more discriminative multimodal representations. The multi-attention module is consist of self-attention, cross-attention and triple-attention components, we employ it to highlight important sentimental representations over time and modality. Especially, we perform the cross-attention component under a "non-alignment" modality data setting to make our method can capture sequence-level interactions between modalities and have a much better multimodal fusion ability (*e.g.*, text → audio) [11]. We also perform self-attention to highlight import elements in each modality, and triple-attention to highlight important modality.

We conduct detailed experiments on the video-level sentiment analysis dataset MOSI [8] and MOSEI [6]. Our method improves F1-score of the state-of-the-art method Self-MM [15] by **3.6%** on MOSI and **2.9%** on MOSEI datasets respectively. We also perform quantitative and qualitative analysis on the test set of both datasets, and the results suggest that *VAE-AMDT* is capable of reducing distance difference among unimodal representations, and fused multimodal representation is discriminative for improving the performance of video-level sentiment analysis.

II. RELATED WORK

A. UNIMODAL SENTIMENT ANALYSIS

Sentiment analysis from people's facial expressions, voices and speech texts have some impressive progress by employing deep learning techniques [1]. Convolutional neural networks (CNN) are employed to do facial expressions recognition (FER) [19], [20]; Recurrent neural networks (RNN) are employed to do speech emotion recognition (SER) [21]–[24]; Language models (*e.g.*, BERT [13]) are finetuned to do textual sentiment analysis [25]–[27]; All these methods focus on learning effective latent representations from single modality. However single modality is not enough to provide comprehensive information to analyze people's complex sentiments. In contrast, our method focus on how to fuse these unimodal latent representations to further improve the performance of sentiment analysis.

B. MULTIMODAL FUSION

Recent works on video-level sentiment analysis are increasing, and aim to gain more effective multimodal representations from various modalities. Several recent works [7], [9]–[11] employ attention mechanism to fuse multimodal representations through modeling interactions across various modalities. Zadeh *et al.* [6] propose a dynamic fusion graph to do inter-multimodal fusion and Wang *et al.* [10] dynamically adjust word representations using its aligned facial expressions and voice representations. However, these methods work with the forced alignment data setting, and are limited to build sequence-level interactions between modality. Our method works with non-alignment data setting, so we can use cross-attention to build sequence-level optimal interactions cross modality.

To further improve the performance of multimodal fusion, recent works [15], [16] focus on how to reduce distance difference of unimodal representations since it is hard for systems to learn a common classifier from various modality domains as shown in Fig. 1. Motivated by adversarial training [29], [30], Mai *et al.* [16] introduce adversarial encoder-decoder-classifier framework to transfer unimodal representations to a joint embedding space, and Yu *et al.* [15] designs an unimodal label auto generation module to better learn unimodal representations so that the distance difference between modality can be reduced. However, their efficacy is limited on the small and imbalance sentiment dataset. We perform adversarial training by using VAE-encoded unimodal representations to better reduce distance difference of unimodal representations.

III. PROBLEM STATEMENT

In this paper, we aim to predict people's sentiment intensity with a given video clip. The video clip includes multimodal

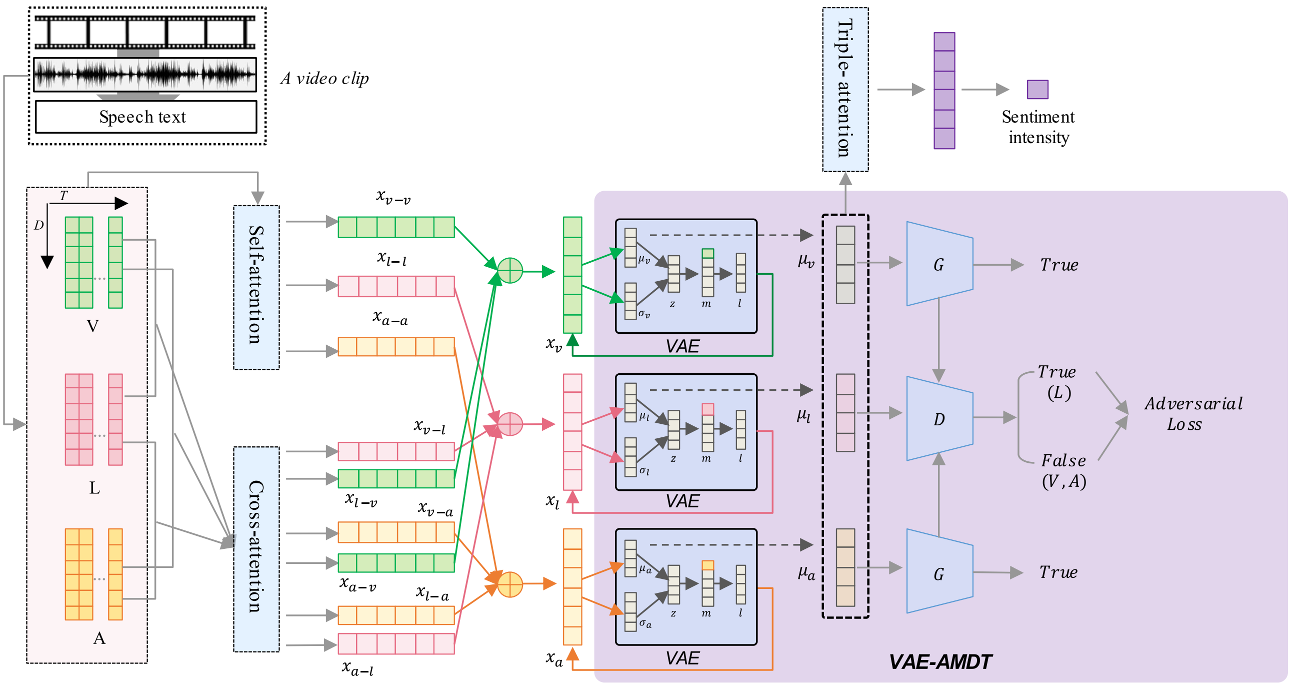


FIGURE 2. Overview of our method: we first perform self-attention (§ V-A1) and cross-attention (§ V-A2) using preprocessed sequence features V, L and A, and then we perform *VAE-AMDT* that consisting of three VAEs and two generators G and one discriminator D to reduce distance difference between unimodal representations (§ V-B). Finally, we use the encoded unimodal representations as the input of triple-attention (§ V-A3) to output one sentiment intensity result. Here, unimodal representations x_v, x_l and x_a indicate concatenations of the output of attention layers for each modality. μ_v, μ_l and μ_a are encoded unimodal representations with *VAE-AMDT*.

signals: people's face image frames (I_v), audio (I_a) and speech text (I_t). We regard this task as a regression task, and our model takes I_v, I_a and I_t as inputs and outputs one sentiment intensity $y \in R$. Here, R is in the range of $[-3, 3]$.

IV. MODALITY DATA PREPROCESSING

Given a video clip, we first drop out data that does not contain all of I_v, I_a and I_t to ensure our model works properly, and then we process each unimodal signal following below techniques to obtain their sequence features:

1) For the visual modality, we first use OpenFace [31] to extract I_v, and then we initialize visual sequence features $V \in \mathbb{R}^{T_v \times D_v}$ by encoding facial expression representations from I_v using a pretrained FER model [33]. Here, the FER model is pretrained on the VGG-Face dataset [34]. Given an extracted face image, we perform that pretrained FER model and use its prediction result as facial expression representations. The facial expression result is represented with a 8-dimensional vector. The More details in Albanie's website[1].
2) For the linguistic modality, we initialize language sequence features $L \in \mathbb{R}^{T_l \times D_l}$ by extracting sentence embeddings of I_t using a pretraining language model RoBERTa [28].
3) For the acoustic modality, we initialize audio sequence features $A \in \mathbb{R}^{T_a \times D_a}$ by extracting log-mel filter banks from I_a [22].

In this paper, to solve one problem of different video clip lengths, we do padding and truncation to adjust the length of V, L and A respectively. We set T_v, T_l and T_a to 64, 100 and 128, and D_v, D_l and D_a to 8, 1024 and 128.

V. METHODOLOGY

In this section, we explain our method in detail. As shown in Fig. 2, our method includes *VAE-AMDT* and a multi-attention module that consists of self-attention, cross-attention and triple-attention components. We jointly train *VAE-AMDT* and the multi-attention module to reduce the distance difference between unimodal representations and fuse multimodal representations to do sentiment intensity prediction.

A. MULTI-ATTENTION MODULE

1) SELF-ATTENTION

The self-attention is designed to highlight key sequence elements [12], [13], and performed by taking V, A and L as inputs and output self-attention vector $x_{(v\to v)}$, $x_{(l\to l)}$ and $x_{(a\to a)}$, as follows:

$$X_{(m)} = f_m(X) \tag{1}$$

$$\alpha_{(m\to m)} = \text{softmax}(X_{(m)} \cdot X_{(m)}^T) \tag{2}$$

$$x_{(m\to m)} = f_s\left(\frac{\sum_{t=1}^{T_m} \alpha_{(m\to m)} \cdot X_{(m)}}{T_{(m)}}\right) \tag{3}$$

where $f_m : \mathbb{R}^{T_m \times D_m} \to \mathbb{R}^{T_m \times D}$ is a linear transformation. We perform f_m with $X \in \{V, L, A\}$ to output $X_{(m)}, m \in \{v, a, l\}$ and they have a same dimension D. We then calculate

[1] https://www.robots.ox.ac.uk/ albanie/mcn-models.html

attention weight $\alpha_{(m\rightarrow m)}$ and get self-attention vector $x_{(m\rightarrow m)}$ via a 2-layer MLP $f_s : \mathbb{R}^D \rightarrow \mathbb{R}^D$.

2) CROSS-ATTENTION

We perform cross-attention between any two modalities to highlight correlated sequence elements over modality. For example, corresponding to one speech text "I enjoyed the party today"., the word "enjoy" should attend to the enjoyable facial expressions, and its cross-attention weight $\alpha_{(m1\rightarrow m2)}$ should be learned with a high score. We use $m1$ and $m2$ to indicate different modality. We perform cross-attention in two attentional directions to get cross-attention vector $x_{(m1\rightarrow m2)}$ and $x_{(m2\rightarrow m1)}$, as follows:

$$\alpha_{(m1\rightarrow m2)} = \text{softmax}(X_{(m1)} \cdot X^T_{(m2)}) \tag{4}$$

$$x_{(m1\rightarrow m2)} = f_s\left(\frac{\sum_{t=1}^{T_{m2}} \alpha_{(m1\rightarrow m2)} \cdot X_{(m2)}}{T_{(m2)}}\right) \tag{5}$$

As shown in Fig. 2, we concatenate self-attention and cross-attention vectors for each modality to get unimodal representations x_m, as follows:

$$x_v = x_{(v\rightarrow v)}||x_{(l\rightarrow v)}||x_{(a\rightarrow v)} \tag{6}$$

$$x_l = x_{(l\rightarrow l)}||x_{(v\rightarrow l)}||x_{(a\rightarrow l)} \tag{7}$$

$$x_a = x_{(a\rightarrow a)}||x_{(l\rightarrow a)}||x_{(v\rightarrow a)} \tag{8}$$

where "||" is the concatenation operation. We take x_v, x_l and x_a as inputs of *VAE-AMDT* (§ V-B).

3) TRIPLE-ATTENTION

We fuse *VAE-AMDT* encoded unimodal representaions μ_v, μ_l and μ_a by using triple-attention so that the important unimodal representaions can be highlighted. We stack μ_v, μ_l and μ_a in a list and then perform Eqs. (2) and (3) to get a multimodal representation vector x. Finally, we perform linear regression for sentiment intensity prediction by employing mean squared error (MSE) loss function $\mathcal{L}_m$, as follows:

$$\mathcal{L}_m(\mathbf{y}, \widehat{\mathbf{y}}) = \frac{1}{n}\sum_{i=1}^{n} |f_r(x_i) - \widehat{y}_i|^2 \tag{9}$$

where $f_r : \mathbb{R}^D \rightarrow \mathbb{R}^1$ is a linear transformation, used to output one sentiment intensity result. n represents the size of data batch and $\widehat{\mathbf{y}}$ is ground truth label.

B. VAE-AMDT

VAE-AMDT is composed of three VAEs and two generators G and one discriminator D (Fig. 2). We jointly train it with the multi-attention module to transfer x_v, x_l and x_a to a joint embedding space and use its output μ_v, μ_l and μ_a to predict sentiment intensity (§ V-A3). We show how to learn VAEs and how G and D worked in the adversarial training process as follows:

1) VARIATIONAL AUTO-ENCODER (VAE)

The Kullback-Leibler Divergence (KLD) term of VEA allows us to regularize the encoder to produce a latent vector z that follows a standard normal distribution [4], [5]. As a result, we have each mean layer $\mu_{(\mathbf{m})}$ that follows a similar distribution [5]. To further include modality type information in the encoder, we define a one-hot vector to represent modality types and concatenate it with the decoder vector m for each modality. Following a MLP layer l, we maximize the loss function $\mathcal{L}_{vae}$ to learn VAEs together as follows [4]:

$$\begin{aligned}\mathcal{L}_{vae}(\theta, \phi) = \sum_{r=1}^{R}\sum_{n=1}^{N}\{-\beta KL(Q_\phi(\mathbf{z}|\mathbf{x}_n^r)||P_\theta(\mathbf{z}))\\ + \mathbb{E}_{Q_\phi(\mathbf{z}|\mathbf{x}_n^r)}\left[\log P_\theta(\mathbf{x}_n^r|\mathbf{z}, \mathbf{m})\right]\}\end{aligned} \tag{10}$$

where ϕ and θ denote encoder and decoder parameters respectively. R denotes the number of modalities and N denotes the data size. We set β to 0.5. This is a trade-off coefficient that allows the model prioritize one term over the other. KL represents KLD term, used to constrain the variational posterior $Q_\phi(\mathbf{z}|\mathbf{x})$ close to the prior $P_\theta(\mathbf{z})$. The second term on the right-hand side of Eq.10 indicates the values of the expected log-likelihood generated by the decoder P_θ. To maximize it to enforce $\mathbf{z}$ return to the original data space with the constraint $\mathbf{m}$. Here, $\mathbf{m}$ is the modality type vector. When KL is minimized, the encoder Q_ϕ is also constrained by $\mathbf{m}$. As a result, the modality type information can affect the encoder optimization and to make the encoder represent modality type information as well.

2) ADVERSARIAL TRAINING

We take VAE encoded unimodal representations μ_v, μ_l and μ_a as the input. To further reduce the distance between any two unimodal representations, we introduce two G to generate fake linguistic modal representations from visual and acoustic modality and then design a D to discriminate the real linguistic modal representation from generated fake representations by employing an adversarial loss $\mathcal{L}_{at}$. In addition, we perform binary classification for the generator by using binary cross entropy loss (BCELoss). We jointly train two G and one D to as follows:

$$\begin{aligned}\mathcal{L}_m^G &= arg\min_{E_m} V(E_m),\\ V(E_m) &= \mathbb{E}_{\mu_\mathbf{m}\sim Q_{\psi_m}(\mu_\mathbf{m})}\left[\log E(\mu_m)\right]\\ &\quad + \mathbb{E}_{\mu_m\sim Q_{\psi_m}(\mu_m)}\left[\log\left(1 - E_m(\mu_m)\right))\right]\end{aligned} \tag{11}$$

where E_m indicates generator of modality $m \in \{v, a\}$.

$$\begin{aligned}\mathcal{L}^D &= arg\min_{E_v,E_a}\max_{D} V(E_v, E_a, D),\\ V(E_v, E_a, D) &= \mathbb{E}_{\mu_\mathbf{l}\sim Q_{\psi_l}(\mu_\mathbf{l})}\left[\log D(\mu_\mathbf{l})\right]\\ &\quad + \mathbb{E}_{\mu_{(v)}\sim Q_{\psi_v}(\mu_{(v)})}\left[\log\left(1 - D\left(E_v(\mu_{(v)})\right)\right)\right]\\ &\quad + \mathbb{E}_{\mu_{(a)}\sim Q_{\psi_a}(\mu_{(a)})}\left[\log\left(1 - D\left(E_a(\mu_{(a)})\right)\right)\right]\end{aligned} \tag{12}$$

Consequently, we have $\mathcal{L}_{at}$ for adversarial training.

$$\mathcal{L}_{at} = \mathcal{L}_v^G + \mathcal{L}_a^G + \mathcal{L}^D \tag{13}$$

TABLE 1. The size of dataset.

Dataset	Train	Validation	Test	Total
MOSI	1257	229	686	2172
MOSEI	9473	1206	2710	13389

C. LEARNING

We finally have a joint loss $\mathcal{L}$ for training the multi-attention module and VAE-AMDT, as follows:

$$\mathcal{L} = \alpha\mathcal{L}_m + \beta\mathcal{L}_{ave} + \gamma\mathcal{L}_{at} \tag{14}$$

where α, β, γ are hyperparameters, which are used to indicate the importance of each loss value. We empirically set them as 1.

VI. EXPERIMENT

A. DATASET

We evaluate our method on using video-level sentiment analysis dataset MOSI [8] and MOSEI [6]. Both datasets are collected from online video: MOSI contains 2,199 opinion video clips and MOSEI contains more than 65 hours video from more than 1000 speakers and 250 topics. To ensure our method behaves correctly, we drop out data that does not contain all of modalities. Tab. 1 shows the number of data in both datasets in detail. MOSEI dataset is over 6x larger than MOSI dataset. Both datasets are annotated in the range of the [−3,3] Likert scale, *i.e.*, [-3: highly negative, −2: negative, −1: weakly negative, 0: neutral, +1: weakly positive, +2: positive, +3: highly positive]. From the data distribution over annotations in Fig. 3, we have very imbalanced data annotations for both datasets. Especially, there is over 65% of MOSEI dataset are annotated in the range of [−1, 1].

B. METRIC

We use the mean absolute error (MAE), accuracy (A^2) and weight $F1$ score as evaluation metric. A^2 is a binary accuracy metric, the prediction result $y < 0$ are belonged to "Negative" class and $y \geq 0$ are belonged to "Non-negative" class (Fig. 3). Furthermore, due to the small and imbalanced dataset, we also use precision-recall curve to show the model's performance at various threshold settings.

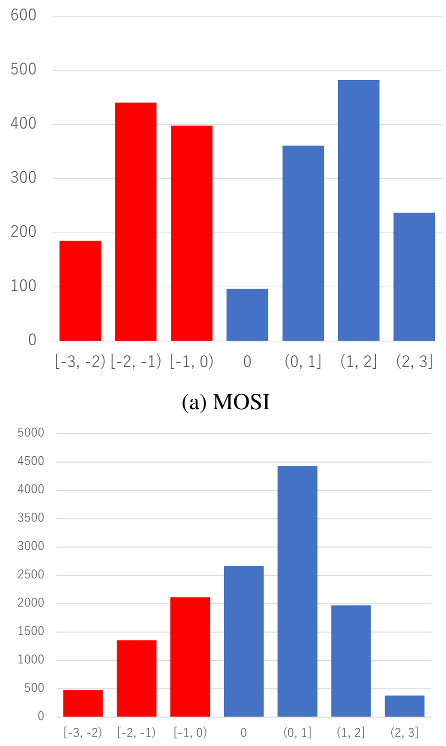
C. FULL MODEL HYPERPARAMETERS

We show full hyperparameters of our model on MOSI and MOSEI dataset in Tab. 2. We use AdamW [35] as our optimizer, with ϵ = 1e-8. We use cosine annealing schedular [36] to adjust the learning rate (1e-8). We also show the feature size of each attention component in our multi-attention module (Fig. 2) in details. Our hidden layer size (f_m) is different from datasets, so that we have different hyperparameters for training their best performance (Tab. 2: "Training").

D. PERFORMANCE

As shown in Tab. 3, under the same modality alignment setting (non-alignment), our method achieves much lower MAE result than Self-MM(+) by over ***0.16*** (MOSI) and ***0.05*** (MOSEI). Especially, a low MAE indicates that our method is superior on the sentiment regression problem. Compared to Self-MM(+), we also note that our method improves MAE results better with MOSI than with MOSEI. This suggests that VAE-AMDT is much effective for relatively small datasets (Tab. 1). Here, Self-MM(+) is trained by using the same preprocessed data in our method (§ IV). Especially, we use a same pretrained RoBERTa model to encode speech text for a fair comparison between Self-MM(+) and our approach. We take out the data that lacks some modalities so that we can fairly compare their performance in terms of the modal fusion capability. To fairly confirm the binary classification ability of models trained with imbalanced annotations (Fig. 3), in addition to the accuracy (A^2) comparisons, we also show precision-recall curve for both MOSI and MOSEI in Fig. 4. The results suggest that our method is superior to Self-MM(+). Even though Self-MM(+)'s A^2 result (84.6%) is higher than our method (82.8%), when both precision and recall scores are over 80% as shown in the precision-recall curve graph (Fig. 4b), our method is still better than Self-MM(+). We also show the result of our method trained in a 10-fold cross-validation strategy (CV), which is a bit worse due to the small and imbalanced dataset, but it is still better than Self-MM(+) except A^2 for MOSEI. This result also suggests that our method is not overfitting to the training set.

We additionally compare the number of parameters of Self-MM and our method. Self-MM finetunes the pretrained BERT model [13], so it needs to reuse and update BERT's



(a) MOSI

(b) MOSEI

FIGURE 3. Annotation distributions on MOSI (a) and MOSEI (b). We show "negative" classes in red color and "non-negative" classes in blue color.

TABLE 2. Full hyperparameters for our model. "dim" indicates the number of dimensions. "Self", "Cross" and "Triple" indicate self-attention, cross-attention and triple-attention components respectively.

		MOSI	MOSEI
Audio	Sample rate	44.1KHz	
	FFT hop length	0.02s	
	FFT window size	0.01s	
	Mel bins	128	
	Sequence length	128	
Image	Frame rate	8fps	
	Face detection	OpenFace [2]	
	Face frame size	128*128	
	Facial expressions feature (dim)	8	
	Sequence length	64	
Text	Tokenization	Roberta Tokenization [3]	
	Embeddings(dim)	1024	
	Sequence length	100	
f_m	Feature size (input)	V:(B, 64, 8); L:(B, 100, 1024); A:(B, 128, 128)	V:(B, 64, 8); L:(B, 100, 1024); A:(B, 128, 128)
	Feature size (output)	V:(B, 64, 128); L:(B, 100, 128); A:(B, 128, 128)	V:(B, 64, 64); L:(B, 100, 64); A:(B, 128, 64)
Self	Feature size (input)	V:(B, 64, 128); L:(B, 100, 128); A:(B, 128, 128)	V:(B, 64, 64); L:(B, 100, 64); A:(B, 128, 64)
	Feature size (output)	V:(B,128); L:(B, 128); A:(B, 128)	V:(B,64); L:(B, 64); A:(B, 64)
Cross	$[v \rightleftharpoons l]$ Feature size (input)	V:(B, 64, 128); L:(B, 100, 128)	V:(B, 64, 64); L:(B, 100, 64)
	$[v \rightleftharpoons l]$ Feature size (output)	$[v \rightarrow l]$ and $[l \rightarrow v]$:(B, 128)	$[v \rightarrow l]$ and $[l \rightarrow v]$:(B, 64)
	$[v \rightleftharpoons a]$ Feature size (input)	V:(B, 64, 128); A:(B, 128, 128)	V:(B, 64, 64); A:(B, 128, 64)
	$[v \rightleftharpoons a]$ Feature size (output)	$[v \rightarrow a]$ and $[a \rightarrow v]$:(B, 128)	$[v \rightarrow a]$ and $[a \rightarrow v]$:(B, 64)
	$[a \rightleftharpoons l]$ Feature size (input)	A:(B, 128, 128); L:(B, 100, 128)	A:(B, 128, 64); L:(B, 100, 64)
	$[a \rightleftharpoons l]$ Feature size (ouput)	$[a \rightarrow l]$ and $[l \rightarrow a]$:(B, 128)	$[a \rightarrow l]$ and $[l \rightarrow a]$:(B, 64)
Joint	Feature size (input)	x_v, x_l and x_a:(B, 384)	x_v, x_l and x_a:(B, 192)
Triple	Feature size (input)	(B,3,32) $[\mu_v, \mu_l, \mu_a]$	
	Feature size (output)	(B,32) $[x]$	
Optimizer	Peak learning rate	1e-4	
	Weight decay	0	
	AdamW β	0.9	
	AdamW ϵ	1e-8	
	Schedular	CosineAnnealingLR	
Training	Loss function	Mean Squared Error (MSE)	
	GPU	GTX 1080 Ti	
	Batch size	4	20
	Training epochs	200	80
	Parameters	3.3M	1.7M
	Training time	1h13m	46m
	Inference time	0.000738	0.000125
	Training time (Self-MM)	3h29m	-
	Inference time (Self-MM)	0.001131	-

parameters, and the training parameters exceed ***100M***. This is ***33X*** larger than our method ***(3.3M)***. Since we utilize the pretrained RoBERTa [28] to embed speech text during preprocessing (§ IV), it is not essential to update massive pretrained parameters. As a result, we can not only train our method in a short time ($\frac{1}{3}$th of Self-MM) as shown in Tab. 2, but also achieves a model that is ***1.5X*** faster than Self-MM for inference.

E. EFFECT OF VAE-AMDT

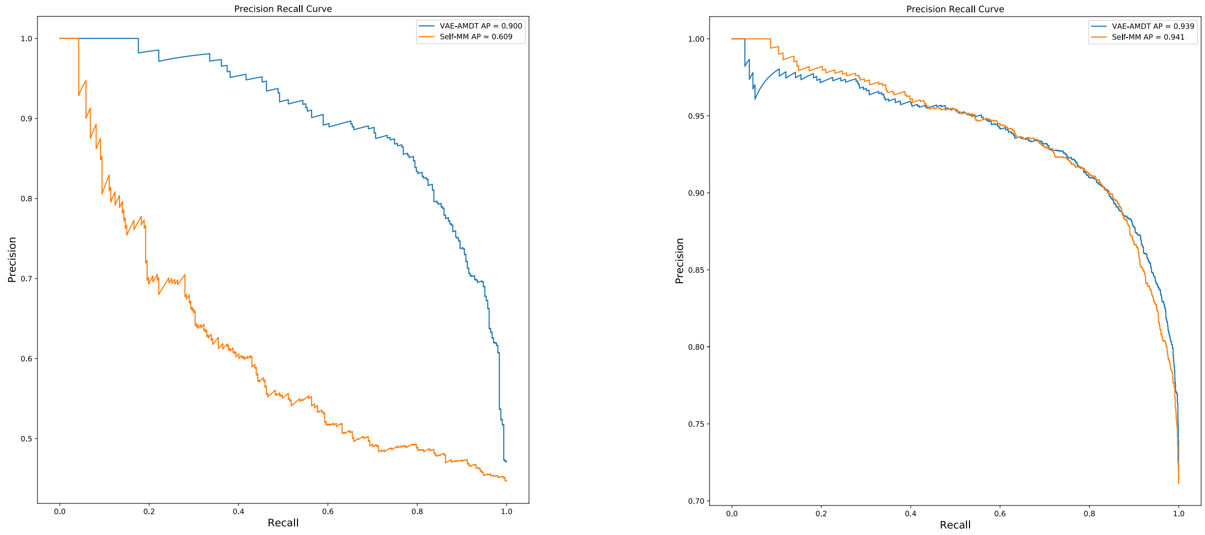
We first show the comparison results of our method built (w/o and w/) *VAE-AMDT* in Tab 4. The results suggest that our proposed *VAE-AMDT* is effective for improving the performance of model only built by employing the multi-attention module (§ V-A). We further study the effect of *VAE-AMDT* through quantitative and qualitative analysis.

1) MAXIMUM MEAN DISCREPANCY SCORE(MMD)

We do quantitative analysis by analyzing maximum mean discrepancy (MMD) on both MOSI and MOSEI test sets. The MMD is a kernel-based approach that is used to measure the distance between two probability distributions [32]. We use encoded unimodal representations μ_v, μ_l and μ_a to

TABLE 3. Comparison of *VAE-AMDT* and state-of-the-art results in both MOSI and MOSEI. *VAE-AMDT* outperforms state-of-the-art Self-MM (*MAE/F1*) by over 0.16/3.6 point (MOSI) and 0.05/2.9 point (MOSEI). Here, the lower the MAE, the better the performance. (*) indicates that the results are referenced from the Self-MM paper. (+) indicates that Self-MM is trained by using the same preprocessed data in our method; (CV) indicates the result of the 10-fold cross validation.

Model	MOSI			MOSEI			Modality alignment
	MAE	A^2	F1	MAE	A^2	F1	
Graph-MFN [7]	0.965	77.4	77.3	-	76.0	76.0	Yes
RAVEN [10]	0.915	78.0	76.6	0.614	79.1	79.5	Yes
ARGF [16]	-	81.3	81.5	-	-	-	Yes
MulT [17]	0.861	81.5	80.6	0.580	-	-	Yes
MISA (*) [18]	0.804	80.8	80.8	0.568	82.6	82.7	Yes
MAG-BERT (*) [14]	0.731	82.5	82.6	0.539	83.8	83.7	Yes
Self-MM (*) [15]	0.713	84.0	84.4	0.530	82.8	82.5	No
Self-MM (+) [15]	0.885	80.6	80.6	0.579	**84.6**	84.6	No
VAE-AMDT	**0.716**	**84.3**	**84.2**	**0.526**	82.8	**87.5**	No
VAE-AMDT (CV)	0.745	82.2	82.2	0.529	81.6	86.2	No
Human [7]	0.710	85.7	87.5	-	-	-	No



(a) MOSI

(b) MOSEI

FIGURE 4. The precision-recall curve is created by using VAE-AMDT's test prediction results on both datasets. The curve indicates that VAE-AMDT outperforms Self-MM when both precision and recall scores exceed 0.8. Here, a better model should perform better for both metrics.

TABLE 4. Comparison results of the model trained w/o and w/ VAE-AMDT. The model trained with VAE-AMDT further improves F1 score of (w/o VAE-AMDT) by 3.6% (MOSI) and 1.7% (MOSEI).

Model	MOSI			MOSEI			Modality alignment
	MAE	A^2	F1	MAE	A^2	F1	
w/o VAE-AMDT	0.808	80.3	80.6	0.603	81.8	85.8	No
w/ VAE-AMDT	**0.716**	**84.3**	**84.2**	**0.526**	**82.8**	**87.5**	No

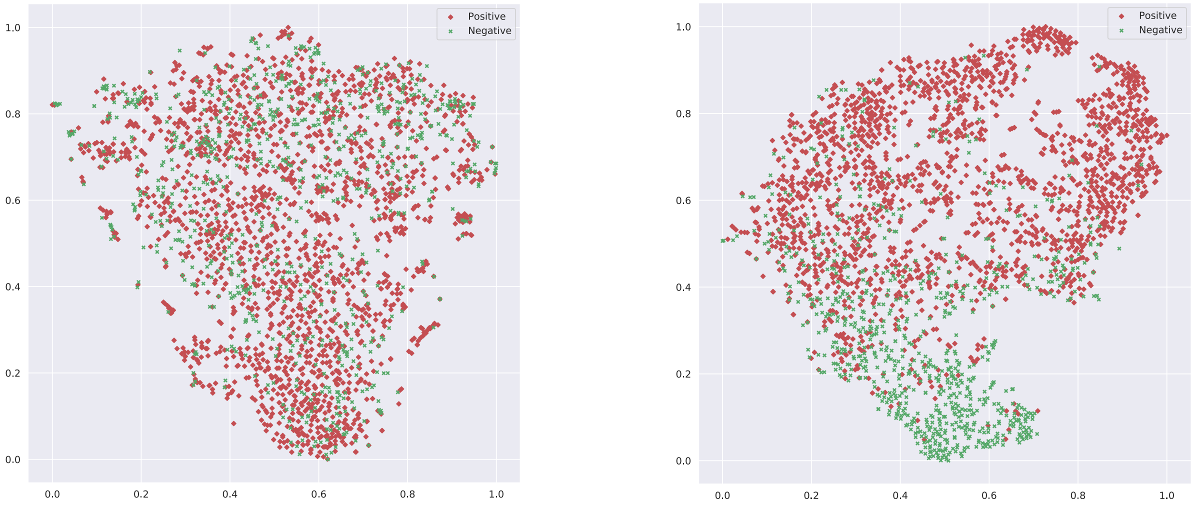
calculate MMD score between any two modality (§ V-B), and show their results in Tab. 5. Our proposed *VAE-AMDT* not only can balance the distance difference between any modality pairs (*e.g.*, v→l, a→l and v→a), but also reduce their average distance difference in total and prove the efficacy of *VAE-AMDT*.

2) VISUALIZATION

To further explain the efficacy of *VAE-AMDT*, we perform qualitative analysis by visualizing the encoded unimodal representations using t-SNE, and show the result on MOSEI test set in Fig. 5. We concatenate encoded unimodal representations μ_v, μ_l and μ_a, and use t-SNE to map them into a joint

TABLE 5. MMD results show that not only can the model (w/ VAE-AMDT) balance the distance between any two modality, but the average result is lower than the model (w/o VAE-AMDT).

Method	MOSI				MOSEI			
	$v \to l$	$a \to l$	$v \to a$	$Average$	$v \to l$	$a \to l$	$v \to a$	$Average$
w/o VAE-AMDT	**0.51**	**0.17**	1.44	0.71	0.98	0.92	0.45	0.79
w/ VAE-AMDT	0.68	0.53	**0.33**	**0.51**	**0.28**	**0.30**	**0.25**	**0.28**

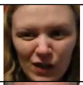
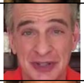
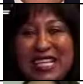
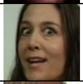


(a) w/o VAE-AMDT

(b) w/ VAE-AMDT

FIGURE 5. Visualization result on MOSEI. The green color indicates "negative" class and the red color indicates "positive" (include "neutral") class. The model (w/ VAE-AMDT) classifies both classes by discriminative representations.

TABLE 6. Case study on MOSEI test set. The predicted sentiment intensity by our method is close to ground truth.

ID	Speech text	Face image	Sentiment intensity	
			Ground Truth	Prediction
1	This movie (umm) if you saw previews for it it looks kind of funny, but this movie actually wasn't very funny		-1.33	-1.69
2	On the other hand he's battling against his fellow atheists who deny that there are any objective moral values and duties		0.00	0.13
3	Millions of women, men, and children have better lives today thanks to the work that many of you have done for decades.		2.00	1.46
4	We will accomplish these goals by reviewing the law that pertains to mandatory child abuse reporting.		1.00	0.82
5	Get ready for this to be bad." That's not good, that's not what you want to do.		1.00	0.52
6	What I cared about was my finances and I felt like I was a pretty smart person and yet you know my financial life was not going where I wanted it to go.		-1.67	-1.27
7	I'm pleased that the United States is represented in Doha by Attorney General-Eric Holder and one of my key White House advisors, Mike Froman.		1.67	0.84

embedding space. By applying *VAE-AMDT* (Fig. 5b), the dots indicating "negative" and "non-negative" classes tend to split into two clusters and prove that *VAE-AMDT* is capable of obtaining discriminative multimodal representations.

TABLE 7. Ablation study of the multi-attention module on MOSEI dataset. All models are trained w/o VAE-ADMT, (self, cross, triple)-attention shows the lowest MAE score compared to others.

Attention type	Sentiment intensity		
	MAE	A^2	$F1$
self-attention	0.688	80.2	85.5
(self, cross)-attention	0.683	81.2	**86.3**
(self, cross, triple)-attention	**0.603**	**81.8**	85.8

TABLE 8. Effect of modality. The test results show that adding modality improves performance.

Modality	MOSI			MOSEI		
	MAE	A^2	$F1$	MAE	A^2	$F1$
Image	1.467	43.4	57.3	0.854	70.0	82.4
Audio	1.498	46.9	54.7	0.867	70.0	83.4
Text	0.990	83.8	75.7	0.698	78.9	84.3
Image, Audio	1.473	52.0	56.2	0.837	67.9	78.6
Audio,Text	0.875	80.0	69.0	0.646	82.8	**87.9**
Image,Text	1.140	74.9	68.1	0.593	82.7	87.8
Image,Text,Audio	**0.716**	**84.3**	**84.2**	**0.526**	**82.8**	87.5

F. ABLATION STUDY

To prove the efficacy of all components in our method, we study the Multi-attention module and Modality respectively. Here, we discuss all comparison results based on the MAE metric. Since we consider that the MAE metric should be more reliable than the A2 and F1 metric on the regression learning, especially for the small and imbalanced datasets.

1) MULTI-ATTENTION MODULE

To confirm the effect of all components of the multi-attention module, we show the comparison results of the model trained by employing different attention component in Tab. 5. For the model employing triple-attention w/o *VAE-AMDT*, we use unimodal representations x_v, x_l and x_a instead of μ_v, μ_l and μ_a. The result suggests that (self, cross, triple)-attention improves performance when use them together. Especially, the MAE result is improved much after adding triple-attention, and suggests its efficacy that highlighting the important modality.

2) MODALITY

To ensure that increasing the number of modality can improve performance, we compare the models that are trained given various modalities as the input, and show the results in Tab. 8. It is clear that adding modality improves performance. However, we note that speech text perform better than other modality (*e.g.*, image, audio). We believe that the language encoder (RoBERTa model [28]) we used is more powerful than encoders used for image and audio.

G. CASE STUDY

We show some data samples from MOSEI test set in Tab. 8. The predicted sentiment intensity by our method is close to ground truth. Although we select samples randomly and the result suggests that our method perform stable with these data. Furthermore, we note that some predicted score is more reasonable than ground truth. For example, the sample (ID:5) is predicted to 0.52, which is lower than ground truth. However we note that the speech text actually represents negative sentiment. These results not only prove that our method is not overfitting to the training set, but also suggest that it is robust to practical use.

VII. CONCLUSION

We proposed (*VAE-AMDT*) and jointly train it with a multi-attention module to reduce the distance difference of various unimodal representations. As a result, we obtained discriminative multimodal representations to further improve performance of video-level sentiment analysis. Our method balanced the distance difference between any modality pairs and reduced their average distance in total (§ VI-E). We finally improve F1-score of the state-of-the-art Self-MM by **3.6%** on MOSI and **2.9%** on MOSEI datasets (§ VI-D), and prove the efficacy of our method in obtaining discriminative multimodal representations (§ VI-E2). In the next step, we will explore more effective approach to improving multimodal fusion, and also plan to use more powerful modal encoders to extract unimodal representations such as face identification method proposed in [37].

REFERENCES

[1] P. V. Rouast, M. T. P. Adam, and R. Chiong, ‘‘Deep learning for human affect recognition: Insights and new developments,’’ *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 524–543, Apr. 2021.

[2] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, ‘‘Multimodal machine learning: A survey and taxonomy,’’ *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[3] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, ‘‘Domain-adversarial training of neural networks,’’ *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, 2016.

[4] Y. Tu, M.-W. Mak, and J.-T. Chien, ‘‘Variational domain adversarial learning for speaker verification,’’ in *Proc. Interspeech*, Sep. 2019, pp. 4315–4319.

[5] D. P. Kingma and M. Welling, ‘‘Auto-encoding variational Bayes,’’ 2013, *arXiv:1312.6114*.

[6] 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, 2018, pp. 1–11.

[7] 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*, vol. 32, 2018, no. 1.

[8] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, ‘‘MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos,’’ 2016, *arXiv:1606.06259*.

[9] 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. ACL*, 2019, p. 6558.

[10] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, ‘‘Words can shift: Dynamically adjusting word representations using nonverbal behaviors,’’ in *Proc. AAAI*, 2019, pp. 7216–7223.

[11] Y. Wang, J. Wu, and K. Hoashi, ‘‘Multi-attention fusion network for video-based emotion recognition,’’ in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 595–601.

[12] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186.

[14] W. Rahman, "Integrating multimodal information in large pretrained transformers," in *Proc. ACL*, 2020, p. 2359.

[15] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. AAAI*, 2021, pp. 10790–10797.

[16] S. Mai, H. Hu, and S. Xing, "Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion," in *Proc. AAAI*, 2020, pp. 164–172.

[17] A. Shenoy and A. Sardana, "Multilogue-Net: A context-aware RNN for multi-modal emotion detection and sentiment analysis in conversation," in *Proc. 2nd Grand-Challenge Workshop Multimodal Lang. (Challenge-HML)*, 2020, pp. 19–28.

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

[19] D. H. Kim, W. J. Baddar, J. Jang, and Y. M. Ro, "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition," *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 223–236, Apr. 2019.

[20] M. Li, H. Xu, X. Huang, Z. Song, X. Liu, and X. Li, "Facial expression recognition with identity and emotion joint learning," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 544–550, Apr. 2021.

[21] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 801–804.

[22] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *Proc. Int. Conf. Platform Technol. Service (PlatCon)*, Feb. 2017, pp. 1–5.

[23] P. Song and W. Zheng, "Feature selection based transfer subspace learning for speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 373–382, Jul. 2020.

[24] R. Panda, R. Malheiro, and R. P. Paiva, "Novel audio features for music emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 11, no. 4, pp. 614–626, Oct. 2020.

[25] W. Jiao, H. Yang, I. King, and M. R. Lyu, "HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition," in *Proc. NAACL*, 2019, pp. 397–406.

[26] J. Islam, R. E. Mercer, and L. Xiao, "Multi-channel convolutional neural network for Twitter emotion and sentiment recognition," in *Proc. NAACL*, 2019, pp. 1355–1365.

[27] C. Huang, A. Trabelsi, X. Qin, N. Farruque, L. Mou, and O. Zaiane, "Seq2Emo: A sequence to multi-label emotion classification model," in *Proc. NAACL*, 2021, pp. 4717–4724.

[28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[29] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[30] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," 2015, *arXiv:1511.05644*.

[31] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.

[32] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[33] S. Albanie and A. Vedaldi, "Learning grimaces by watching TV," in *Proc. Brit. Mach. Vis. Conf.*, 2016.

[34] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 1–12.

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

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

[37] N. Samadiani, G. Huang, Y. Hu, and X. Li, "Happy emotion recognition from unconstrained videos using 3D hybrid deep features," *IEEE Access*, vol. 9, pp. 35524–35538, 2021.

YANAN WANG (Student Member, IEEE) received the B.S. degree in engineering from Aoyama Gakuin University, in 2015, and the M.S. degree in engineering from The University of Electro-Communications, Japan, in 2017. He is currently pursuing the Ph.D. degree in engineering with Keio University, Japan. He works as an Associate Research Engineer with KDDI Research in multimodal modeling topics. His research interests include multimodal representation learning, emotion recognition, knowledge graph, and graph representation learning. He is a Student Member of JSAI, a regular member of IEICE, and an Editorial Committee Member of IEICE Human Communication Group.

JIANMING WU received the B.E. degree from Shanghai Jiao Tong University, in 1998, and the M.E. and Ph.D. degrees from Waseda University, in 2002 and 2005, respectively. He has been with KDDI Research Inc., since 2005, where he is currently the Research Manager of the Multi-Model Communication Laboratory. His research interests include NLP dialogue-AI, facial expression recognition, face id recognition, and multimodal emotion detection. He is a member of IEICE and IPSJ, and an Editorial Committee Member of IEICE Human Communication Group.

KAZUAKI FURUMAI received the B.S. and M.E. degrees from Kobe University, Japan, in 2018 and 2020, respectively. He works as an Associate Research Engineer with KDDI Research in multimodal modeling topics. His research interests include multimodal representation learning, neural language processing, knowledge graph, and graph representation learning.

SHINYA WADA received the B.E. and M.E. degrees from Kyushu University, in 2005 and 2007, respectively. He works as a Senior Manager with the Multimodal Modeling Laboratory, KDDI Research, Inc. His research interests include multimodal representation learning, human activity recognition, and time-series analysis. He is a member of IEICE.

SATOSHI KURIHARA (Member, IEEE) received the B.E. and M.E. degrees in computer science and the Ph.D. degree from Keio University, Tokyo, Japan, in 1990, 1992, in 2000, respectively. In 1992, he joined the Basic Research Division, Nippon Telegraph and Telephone Corporation (NTT). In 2004, he joined the Graduate School of Information Science and Technology, Institute of Scientific and Industrial Research, Osaka University, Osaka, Japan. In 2013, he joined the Graduate School of Information Systems, The University of Electro-Communications. Since 2018, he has been with the Faculty of Science and Technology, Keio University, as a Professor. Since April 2021, he has been the Director of the Center of Advanced Research for Human-AI Symbiosis Society. His research interests include multiagent systems, ubiquitous computing, and complex network research. He is a member of ACM, AAAI, the Information Processing Society of Japan (IPSJ), the Japan Society of Artificial Intelligence (JSAI), the Institute of Electronics, Information and Communication Engineers (IEICE), the Society for Economic Science with Heterogeneous Interacting Agents (ESHIA), and the Japan Society of Software Science and Technology (JSSST).