Learning Modality-Specific and -Agnostic Representations for Asynchronous Multimodal Language Sequences

Dingkang Yang
Academy for Engineering and Technology, Fudan University

Haopeng Kuang
Academy for Engineering and Technology, Fudan University

Shuai Huang
Academy for Engineering and Technology, Fudan University
Engineering Research Center of AI and Robotics, Ministry of Education
Artifical Intelligence and Unmanned Systems Engineering Research Center of Jilin Province

Lihua Zhang*
Academy for Engineering and Technology, Fudan University
Jilin Provincial Key Laboratory of Intelligence Science and Engineering
Ji Hua Laboratory
lihuazhang@fudan.edu.cn

ABSTRACT

Understanding human behaviors and intents from videos is a challenging task. Video flows usually involve time-series data from different modalities, such as natural language, facial gestures, and acoustic information. Due to the variable receiving frequency for sequences from each modality, the collected multimodal streams are usually unaligned. For multimodal fusion of asynchronous sequences, the existing methods focus on projecting multiple modalities into a common latent space and learning the hybrid representations, which neglects the diversity of each modality and the commonality across different modalities. Motivated by this observation, we propose a Multimodal Fusion approach for learning modality-Specific and modality-Agnostic representations (MFSA) to refine multimodal representations and leverage the complementarity across different modalities. Specifically, a predictive self-attention module is used to capture reliable contextual dependencies and enhance the unique features over the modality-specific spaces. Meanwhile, we propose a hierarchical cross-modal attention module to explore the correlations between cross-modal elements over the modality-agnostic space. In this case, a double-discriminator strategy is presented to ensure the production of distinct representations in an adversarial manner. Eventually, the modality-specific and -agnostic multimodal representations are used together for downstream tasks. Comprehensive experiments on three multimodal datasets clearly demonstrate the superiority of our approach.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; *Neural networks*; • **Information systems** → *Multimedia streaming*.

* indicates corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '22, October 10–14, 2022, Lisboa, Portugal.
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547755

KEYWORDS

multimodal fusion, representation learning, adversarial learning

ACM Reference Format:
Dingkang Yang, Haopeng Kuang, Shuai Huang, and Lihua Zhang. 2022. Learning Modality-Specific and -Agnostic Representations for Asynchronous Multimodal Language Sequences. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3547755

1 INTRODUCTION

Video flows usually involve time-series data from multiple modalities, such as natural language, visual information, and acoustic behaviors. Analyzing videos from a multimodal perspective can facilitate a superior understanding of human intentions and expressions (*e.g.*, emotions). To fully utilize the rich information and knowledge provided by multiple modalities, the core is to fuse these multimodal sequence data effectively. Several prominent works [1, 8, 31, 32, 40, 49, 53] of multimodal fusion have breathed fresh energy into the multimodal video understanding community, benefiting from the excellent performance achieved by deep learning algorithms [4, 15, 17, 21, 27–29, 41, 42, 50]. Nevertheless, in practice, the collected multimodal data are usually asynchronous due to the variable receiving frequency for sequences of different modalities [38]. For instance, the video frame with a stiff facial expression may relate to the negative voice in the past. The asynchrony across different modalities can increase the difficulty on conducting effective multimodal fusion.

To this end, most previous works [13, 25, 36, 39, 44, 46, 48] tackle the above issue via word-level alignment. Specifically, the visual and acoustic sequences are aligned manually in the resolution of the textual words. Unfortunately, the alignment process usually requires domain-related knowledge engineering and consumes a lot of time and labour. Moreover, the word-level multimodal fusion ignores the long-range dependencies between elements from different modalities. Recent works [26, 30, 38] deal directly with asynchronous multimodal sequences based on the cross-modal attention to progress the development of effective multimodal fusion on the unaligned data. The Multimodal Transformer (MulT) [38] are proposed to reinforce a target modality repeatedly with the

low-level features from another source modality by learning the attention across the two modalities' features. Based on the cross-modal interaction insight, the Progressive Modality Reinforcement (PMR) [30] and Modality-Invariant Cross-modal Attention (MICA) [26] models are presented one after another. The PMR introduces the message hub to explore the three-way interactions across the involved modalities in the context of multimodal fusion from asynchronous multimodal sequences. However, both the MulT and PMR treat the representation of each modality as a whole, neglecting the interference of modality heterogeneity and distribution gap. The MICA performs the cross-modal attention over modality-invariant space where the distribution gap across modalities is bridged. Nevertheless, learning contextual dependencies on the common space is potentially unrefined and ignores the diversity of each modality.

Motivated by the above observations, we propose a Multimodal Fusion approach for learning modality-Specific and -Agnostic representations (MFSA) to refine multimodal representations effectively. The core strategy of the MFSA is to learn different aspects of the multimodal representations by projecting multiple modalities into modality-specific and -agnostic spaces, respectively. For the modality-specific representations, we first propose a predictive self-attention module to effecively enhance the unique features of each modality and learn the contextual dependencies within the modality. After that, three specific encoders implement the projection of the specific representations. For the modality-agnostic representations, we introduce a hierarchical cross-modal attention module to achieve sufficiently cross-modal interactions and capture meaningful element correlations across modalities. Immediately, a shared agnostic encoder achieves the projection of the agnostic representations. In this case, we propose a double-discriminator adversarial strategy to supervise the learning of different representations and parameters. The MFSA not only considers the commonality across multiple modalities, but also captures the specificity of each modality. Our model achieves significant performance improvements on several video understanding datasets by fusing these refined representations. Overall, we make the following three contributions:

- We present MFSA, a novel approach to learning effective multimodal representations in asynchronous sequences with a feature disentanglement perspective. The MFSA depicts the commonality across multiple modalities and the diversity of each modality by learning modality-specific and modality-agnostic representations.
- We introduce two effective modules for progressively reinforcing and refining the distinct representations based on self-attention and cross-modal attention. The qualitative analyses clearly demonstrate the rationality and necessity of the proposed modules.
- Our MFSA outperforms previous state-of-the-art works on three multimodal video understanding datasets. Comprehensive experiments show the superiority of our approach.

2 RELATED WORK

2.1 Multimodal Sequence Fusion

Video understanding requires the fusion of time-series data from multiple modalities, such as language, visual, and acoustic modalities. Most previous works [7, 18, 23] focus on multimodal fusion of static features extracted from video clips, without considering the inherent dependencies between elements in multimodal sequences. However, the multimodal streams are usually asynchronous due to the variable frame rate for sequences of different modalities. To this end, recent works involve a manual step to align the acoustic and visual sequences in the resolution of textual words before training. These works include shared-private representation learning [46], cyclic translation mechanism [36], recurrent multistage fusion [25], nonverbal temporal interaction [44], etc. However, the manual alignment usually requires a huge amount of labor effort and time. Recently, several works make some attempts to fuse information from asynchronous multimodal sequences. Tsail *et al.* [38] propose the cross-modal attention mechanism to learn the inherent correlations across modalities. Lv *et al.* [30] introduce a message hub to obtain the reinforced features of the source modalities. Liang *et al.* [26] advocate learning correlations between elements over the modality-invariant space.

2.2 Multimodal Reprentation Learning

Multimodal representation learning aims to extract meaningful semantic information from heterogeneous modalities [51]. In addition, the consistency and complementarity of multiple modalities should be considered in this learning paradigm [16, 52]. Recently, more advanced neural network architectures have been proposed to learn multimodal representations. Hai *et al.* [12] present two methods for unsupervised learning of joint multimodal representations using sequence-to-sequence models. Sun *et al.* [37] use the deep canonical correlation analysis to combine different individual features. Gwangbeen *et al.* [33] apply the adversarial learning concept to multimodal learning and only use the category information for multimodal embedding. Besides, Wang *et al.* [43] propose a deep variational canonical correlation analysis to disentangle the shared and private information of multimodal data. Furthermore, the domain separation network [2] extracts the effective representations by explicitly modelling the shared and domain-specific private features of source and target domains. More recently, Hazarika *et al.* [6] propose a framework called MISA, which learns the multimodal representations within instances and projects each modality to two distinct subspaces.

3 APPROACH

3.1 Model Overview

In this section, we detail the proposed Multimodal Fusion approach for learning modality-Specific and modality-Agnostic representations (MFSA). The overall architecture of the MFSA is shown in Figure 1. This paper focuses on performing asynchronous multimodal sequence fusion from three primary modalities, *i.e.*, language (L), visual (V), and audio (A) modalities. These sequences are denoted as $\boldsymbol{X}_L \in \mathbb{R}^{T_L \times d_L}$, $\boldsymbol{X}_V \in \mathbb{R}^{T_V \times d_V}$, and $\boldsymbol{X}_A \in \mathbb{R}^{T_A \times d_A}$, respectively, where $T_{(\cdot)}$ is the sequence length and $d_{(\cdot)}$ is the embedding dimension. Firstly, we preprocess the multimodal sequences to obtain the low-level representations $\boldsymbol{Z}_m \in \mathbb{R}^{T_m \times d}$, where $m \in \{L, V, A\}$. Subsequently, two separate branches are introduced to learn distinct representations from different modalities. The first branch aims to use the proposed predictive self-attention module to enhance the

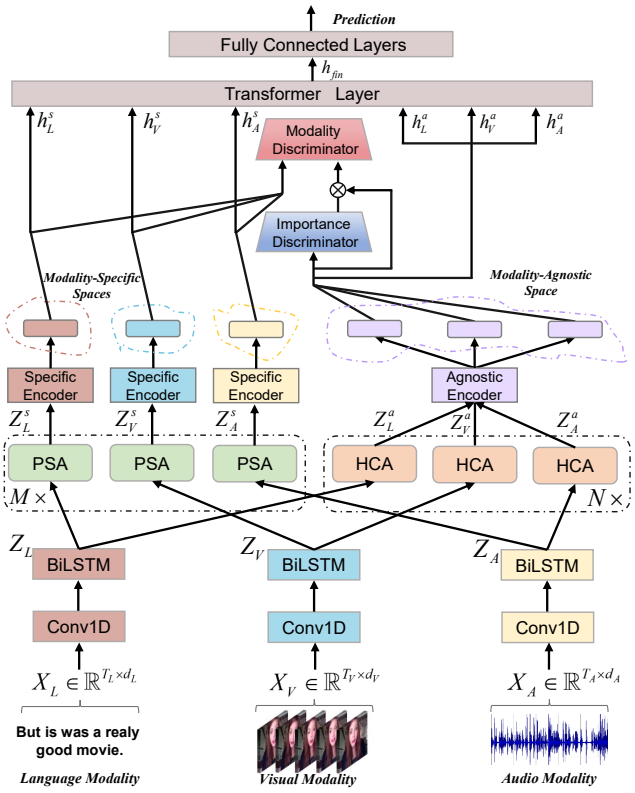


Figure 1: The overall architecture of the proposed model. "PSA" represents a predictive self-attention module. "HCA" represents a hierarchical cross-modal attention module.

features of each modality itself and effectively capture the dependencies with apriori knowledge. After that, we project the enhanced features $\boldsymbol{Z}_m^s \in \mathbb{R}^{T_m \times d}$ into modality-specific spaces via three separate specific encoders to learn the unique characteristics and the diversity of each modality. The second branch focuses on exploring meaningful correlations between elements of different modalities through the proposed hierarchical cross-modal attention module. In this case, a shared agnostic encoder projects the reinforced features $\boldsymbol{Z}_m^a \in \mathbb{R}^{T_m \times d}$ into the modality-agnostic space to learn the commonality and bridge the distribution gap among different modalities. Further, we propose a double-discriminator adversarial strategy to explicitly supervise the production of the modality-specific representations $\boldsymbol{h}_m^s \in \mathbb{R}^{d_h}$ and the modality-agnostic representations $\boldsymbol{h}_m^a \in \mathbb{R}^{d_h}$, and parameter learning of the model. We then concatenate these distinct representations $\boldsymbol{h}_m^s$ and $\boldsymbol{h}_m^a$ as $\boldsymbol{h}_{fin}$, and fuse the features through a vanilla transformer layer. Finally, the refined multimodal representation $\boldsymbol{h}_{fin}$ is fed into the fully connected layers for making predictions.

3.2 Uni-modal Extractor

First, the original multimodal sequences $\boldsymbol{X}_m \in \mathbb{R}^{T_m \times d_m}$ are preprocessed by a 1D temporal convolutional layer and a positional embedding augment opertaion [41], where $m \in \{L, V, A\}$. By controlling the kernel size of the convolutional operation for each modality, the features of different modalities are aligned to the identical dimension represented as $\boldsymbol{X}_m \in \mathbb{R}^{T_m \times d}$. Subsequently, we employ three separate Bi-directional Long Short Term Memory (Bi-LSTM) [15] to obtain the low-level features of multimodal sequences:

$$\boldsymbol{Z}_m = \text{Bi-LSTM}(\boldsymbol{X}_m; \theta_m^{lstm}) \in \mathbb{R}^{T_m \times d}, \tag{1}$$

where θ_m^{lstm} are the network parameters.

3.3 Predictive Self-Attention Module

Transformer [41] is the state-of-the-art for sequential modeling which achieves superior performance. However, as proved by previous work [20], it is difficult for a vanilla attention layer to capture the dependencies effectively without any apriori knowledge. Moreover, the vanilla self-attention mapping of each layer is learned independently, which limits the performance of the sequence representations from different modalities [45]. Based on the above observation and inspiration [45], we introduce a Predictive Self-Attention (PSA) module to capture reliable contextual dependencies and enhance the feature representation of each modality. Specifically, we introduce a convolution-based prediction chain to calculate attention maps for the current module based on the attention map from the previous module. Our insight is that the chain will predict effective attention maps guided by attention patterns from previous modules. Therefore, the self-attention layer in the current PSA module could be dedicated to merging modality-specific knowledge into residual attention maps.

Figure 2.(a) displays the two-layer stacked PSA modules from three modalities for illustration purposes. Following [41], the PSA module contains Querys, Keys, and Values, denoted as $\boldsymbol{Q}_m = LN(\boldsymbol{Z}_m)\boldsymbol{W}_{Q_m}$ with $\boldsymbol{W}_{Q_m} \in \mathbb{R}^{d \times d}$, $\boldsymbol{K}_m = LN(\boldsymbol{Z}_m)\boldsymbol{W}_{K_m}$ with $\boldsymbol{W}_{K_m} \in \mathbb{R}^{d \times d}$, and $\boldsymbol{V}_m = LN(\boldsymbol{Z}_m)\boldsymbol{W}_{V_m}$ with $\boldsymbol{W}_{V_m} \in \mathbb{R}^{d \times d}$, respectively, where $m \in \{L, V, A\}$ and LN stands for layer normalization. We define the matrix of attention logits as $A = \frac{\boldsymbol{Q}_m \boldsymbol{K}_m^T}{\sqrt{d}} \in \mathbb{R}^{T_m \times T_m}$. Assuming there are K heads in the multi-head attention, then we obtain K attention logits maps. These maps construct a tensor $\boldsymbol{A} \in \mathbb{R}^{T_m \times T_m \times K}$, which can be viewed as a $T_m \times T_m$ image with K input channels. In this case, we adopt a 2D convolutional layer with 3×3 kernels to predict the attention maps for the next module. Keeping the output channels are also K, the attention logits maps of all heads can be generated jointly. Immediately, a GeLU [14] activation is utilized to provide non-linearity and sparsity. Finally, the previously predicted attention maps $\text{CNN}(\boldsymbol{A}_{pre})$ are combined with that learned by current dot-product attention $\boldsymbol{A}_{cur}$:

$$\boldsymbol{A} = softmax(\mu \odot \text{CNN}(\boldsymbol{A}_{pre}) + (1 - \mu) \odot softmax(\boldsymbol{A}_{cur})), \tag{2}$$

where $\mu \in [0, 1]$ is a hyper-parameter to balance the importance of two parts. Note that the predictive attention maps are not applied to the PSA module in the first layer. Subsequently, the PSA module computes feed-forwardly as follows:

$$\boldsymbol{Z}_m^s = LN(\boldsymbol{Z}_m) + \boldsymbol{A}\boldsymbol{V}_m, \tag{3}$$

$$\boldsymbol{Z}_m^s = \mathcal{F}_\theta(LN(\boldsymbol{Z}_m^s)) + \boldsymbol{Z}_m^s, \tag{4}$$

where $\mathcal{F}_\theta(\cdot)$ is the position-wise feed-forward network.

In this branch, our goal is to strengthen the pure representation of each modality that serves modality-specific spaces' projection.

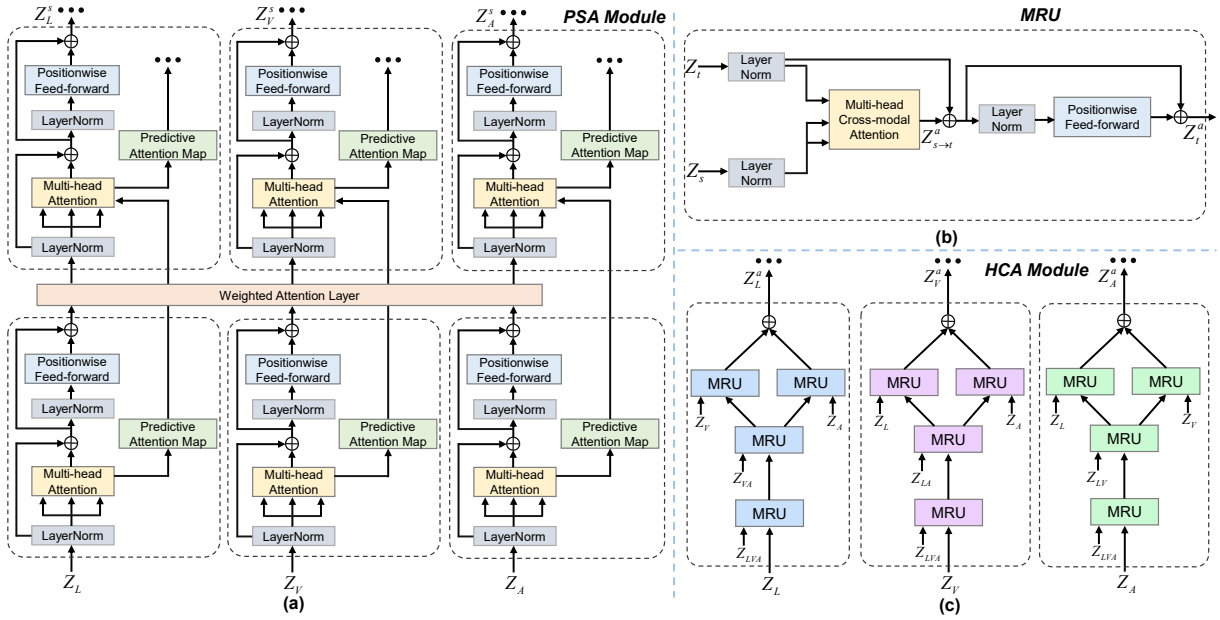


Figure 2: (a) The architecture of the two-layer Predictive Self-Attention (PSA) modules from three modalities. (b) The architecture of a Modality Reinforcement Unit (MRU) in the HCA module. (c) The architecture of the Hierarchical Cross-modal Attention (HCA) modules from three modalities.

However, the heterogeneity across multiple modalities leads to redundant information in multimodal representations [24]. To this end, we present a weighted attention layer after each layer of the PSA module outputs from the three modalities to mitigate the interference of information redundancy. Specifically, we learn adaptive attention weights based on the contribution of each modality to dynamically improve the multimodal representations. For the outputs $\boldsymbol{Z}_m^s \in \mathbb{R}^{T_m \times d}$ from any layer of the PSA module, we first reshape $\boldsymbol{Z}_m^s$ into $\tilde{\boldsymbol{Z}}_m^s \in \mathbb{R}^{T_m \cdot d \times 1}$. Immediately, the dynamic attention weights are calculated as follows:

$$\gamma_m = \boldsymbol{P}_m^T \cdot tanh(\boldsymbol{W}_m \cdot \tilde{\boldsymbol{Z}}_m^s + \boldsymbol{b}_m), \tag{5}$$

$$\psi_m = \frac{exp(\gamma_m)}{\sum_{m \in \{L,V,A\}} exp(\gamma_m)}, \tag{6}$$

where $\boldsymbol{P}_m \in \mathbb{R}^{T_m \cdot d \times 1}$, $\boldsymbol{W}_m \in \mathbb{R}^{T_m \cdot d \times T_m \cdot d}$, and $\boldsymbol{b}_m \in \mathbb{R}^{T_m \cdot d \times 1}$ are the learnable parameters. The weighted multimodal representations are defined as $\boldsymbol{Z}_m^s = \psi_m \odot \boldsymbol{Z}_m^s$. In practice, we stack M-layer PSA modules with the weighted attention layers to enhance the multimodal representations $\boldsymbol{Z}_m^s$ progressively.

3.4 Hierarchical Cross-Modal Attention Module

Although previous studies have achieved significant improvements in cross-modal interactions, the existing models either consider pairwise directional interactions between independent modalities [26, 38], or focus on coarse global interactions [30], resulting in captured cross-modal correlations that could be ambiguous and unreliable. To tackle this issue, we propose a Hierarchical Cross-modal Attention (HCA) module to achieve sufficiently cross-modal interactions and learn effectively cross-modal element correlations. The core idea is to progressively reinforce the representation of the target modality through a potential adaptation process from the source modality $\boldsymbol{Z}_s, s \in \{L, V, A\}$ to the target modality $\boldsymbol{Z}_t, t \in \{L, V, A\}$. We argue that the multimodal representations reinforced by cross-modal interactions have excellent modality adaptability to better serve the modality-agnostic space's projection. More formally, as shown in Figure 2.(c), the HCA module consists of several Modality Reinforcement Units (MRU) to exploit interactions across modalities with different granularities. The multimodal representations are fused in a hierarchical structure and gradually complement each other in a granularity-increasing manner.

Figure 2.(b) illustrates the pipeline of the core unit MRU. Inspired by [41], we embed the target modality as $\boldsymbol{Q}_t = LN(\boldsymbol{Z}_t)\,\boldsymbol{W}_{Q_t}$ with $\boldsymbol{W}_{Q_t} \in \mathbb{R}^{d \times d}$, and the source modality as $\boldsymbol{K}_s = LN(\boldsymbol{Z}_s)\,\boldsymbol{W}_{K_s}$ with $\boldsymbol{W}_{K_s} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{V}_s = LN(\boldsymbol{Z}_s)\,\boldsymbol{W}_{V_s}$ with $\boldsymbol{W}_{V_s} \in \mathbb{R}^{d \times d}$. The cross-modal interaction is denoted as follows:

$$\boldsymbol{Z}_{s \to t}^a = softmax(\frac{\boldsymbol{Q}_t \boldsymbol{K}_s^T}{\sqrt{d}})\boldsymbol{V}_s \in \mathbb{R}^{T_t \times d}. \tag{7}$$

Subsequently, the forward computation is expressed as:

$$\boldsymbol{Z}_t^a = LN(\boldsymbol{Z}_t) + \boldsymbol{Z}_{s \to t}^a, \tag{8}$$

$$\boldsymbol{Z}_t^a = \mathcal{F}_\delta(LN(\boldsymbol{Z}_t^a)) + \boldsymbol{Z}_t^a, \boldsymbol{Z}_t^a \in \mathbb{R}^{T_t \times d}, \tag{9}$$

where $\mathcal{F}_\delta(\cdot)$ is the position-wise feed-forward network. The process of a MRU unit is denoted as $\boldsymbol{Z}_t^a = \text{MRU}(\boldsymbol{Z}_s, \boldsymbol{Z}_t)$. After that, we describe the details of the HCA module with the language modality as the target modality. The low-level representations $\boldsymbol{Z}_m, m \in \{L, V, A\}$ are concatenated with mixed and coarse granularities to obtain $\boldsymbol{Z}_{LVA} = [\boldsymbol{Z}_L, \boldsymbol{Z}_V, \boldsymbol{Z}_A] \in \mathbb{R}^{(T_L+T_V+T_A) \times d}$ and

$Z_{VA} = [Z_V, Z_A] \in \mathbb{R}^{(T_V+T_A)\times d}$, respectively. The hierarchical cross-modal interactions of the HCA module are summarized as:

$$\begin{aligned} \text{Mixed-grained}: \hat{Z}_t^a &= \text{MRU}(Z_{LVA}, Z_L), \\ \text{Coarse-grained}: \check{Z}_t^a &= \text{MRU}(Z_{VA}, \hat{Z}_t^a), \\ \text{Fine-grained}: Z_m^a &= \text{MRU}(Z_V, \check{Z}_t^a) + \text{MRU}(Z_A, \check{Z}_t^a). \end{aligned} \tag{10}$$

In practice, we stack N-layer HCA modules to reinforce the multimodal representations Z_m^a progressively.

3.5 Representation Learning

Modality-Specific and -Agnostic Representations. For learning multimodal representations in asynchronous multimodal sequences, previous approaches either treat each modality as a whole to model, resulting in contextual dependencies that may be unclear [30, 38], or project different modalities into a common latent space to eliminate redundancy, ignoring the diversity of each modality [26]. In comparison, we learn the modality-specific and modality-agnostic representations for each modality to take advantage of the complementary information among multiple modalities. The modality-specific representations focus on the diversity and the unique characteristics of each modality, which are built over the pure multimodal representations Z_m^s. The modality-agnostic representations aim to explore the commonality across different modalities and reduce modality heterogeneity gap, which are built over the refined multimodal representations Z_m^a. To this end, three separate specific encoders and a agnostic encoder are designed to project Z_m^s and Z_m^a into modality-specific and modality-agnostic spaces:

$$\boldsymbol{h}_m^s = \mathcal{S}_m(Z_m^s; \theta_m) \in \mathbb{R}^{d_h}, \tag{11}$$

$$\boldsymbol{h}_m^a = \mathcal{A}(Z_m^a; \theta_{\mathcal{A}}) \in \mathbb{R}^{d_h}, \tag{12}$$

where $\mathcal{S}_m(\cdot; \theta_m)$ denote the specific encoders, which assign separate parameters θ_m for each modality. $\mathcal{A}(\cdot; \theta_{\mathcal{A}})$ denotes the agnostic encoder, which shares the parameters $\theta_{\mathcal{A}}$ across all modalities. These encoders consist of feed-forward neural layers.

Separation Loss. The separation loss aims to encourage the specific and agnostic encoders to produce distinct representations that represent different aspects of the multimodal data. Inspired by domain separation network [2], we adopt soft space orthogonality constraint to penalize redundancy:

$$\mathcal{L}_{sep} = \sum_{m\in\{L,V,A\}} \sum_{i=1}^{n} \| (\boldsymbol{h}_m^s)^T \boldsymbol{h}_m^a \|_F^2, \tag{13}$$

where $\| \cdot \|_F^2$ is the squared Frobenius norm.

Double-Discriminator Adversarial Strategy. To guarantee that $\boldsymbol{h}_m^s$ exactly reflects the unique characteristics of each modality and that $\boldsymbol{h}_m^a$ belongs to a latent space shared across different modalities, we propose a double-discriminator adversarial strategy to identify the modality labels and guide the parameter learning of the specific and agnostic encoders. Our strategy is inspired by applying generative adversarial network [10] to multimodal representation learning [24]. Formally, the ground truth modality labels of $\boldsymbol{h}_m^s$ and $\boldsymbol{h}_m^a$ are denoted as $y_L = [1, 0, 0], y_V = [0, 1, 0], y_A = [0, 0, 1]$, respectively. The importance discriminator is a classifier denoted as $\mathcal{D}_i(\boldsymbol{h}_m^a; \theta_{\mathcal{D}_i}) = softmax((\boldsymbol{h}_m^a)^T \cdot \boldsymbol{W}_i)$, where $\boldsymbol{W}_i \in \mathbb{R}^{d_h\times 3}$ is the weight matrix. Assume that $\mathcal{D}_i(\cdot; \theta_{\mathcal{D}_i})$ has converged to its optimal solution, and the $\boldsymbol{h}_m^a$ belongs to modality m. The importance discriminator gives the likelihood of the $\boldsymbol{h}_m^a$ based on the modality m. If $\mathcal{D}_i(\boldsymbol{h}_m^a; \theta_{\mathcal{D}_i}) \approx 1$, the $\boldsymbol{h}_m^a$ highly involves few modality-agnostic representations across modalities, as it can be fully discriminated from other modalities. Therefore, the degree of the $\boldsymbol{h}_m^a$ as ω_m^a contributing to the modality-agnostic representations should be inversely related to $\mathcal{D}_i(\boldsymbol{h}_m^a; \theta_{\mathcal{D}_i})$ according to $\omega_m^a = 1 - \mathcal{D}_i(\boldsymbol{h}_m^a; \theta_{\mathcal{D}_i})$.

The modality discriminator $\mathcal{D}_m(\boldsymbol{I}; \theta_{\mathcal{D}_m})$ maps the input $\boldsymbol{I}$ to a probability distribution and facilitates the generation of the distinct representations. The input $\boldsymbol{I}$ comes either from the output $\boldsymbol{h}_m^s$ of the specific encoders or from the output $\boldsymbol{h}_m^a$ of the agnostic encoder. After adding the degrees to the modality-agnostic representations for the discriminator $\mathcal{D}_m(\cdot; \theta_{\mathcal{D}_m})$, the agnostic adversarial loss is:

$$\mathcal{L}_{agn} = -\frac{1}{n} \sum_m \sum_{i=1}^{n} (y_m \omega_m^a log(\mathcal{D}_m(\boldsymbol{h}_m^a; \theta_{\mathcal{D}_m}))), \tag{14}$$

where $m \in \{L, V, A\}$. To encourage the modality-specific representations to be projected into different spaces, the modality discriminator is also employed to distinguish the source of the modalities. The specific adversarial loss is as follows:

$$\mathcal{L}_{spe} = -\frac{1}{n} \sum_m \sum_{i=1}^{n} (y_m log(\mathcal{D}_m(\boldsymbol{h}_m^s; \theta_{\mathcal{D}_m}))). \tag{15}$$

3.6 Fusion and Optimization

After obtaining the refined multimodal representations by the adversarial manner, we concatenate all the representations as $\boldsymbol{h}_{fin} = [\boldsymbol{h}_L^s, \boldsymbol{h}_V^s, \boldsymbol{h}_A^s, \boldsymbol{h}_L^a, \boldsymbol{h}_V^a, \boldsymbol{h}_A^a] \in \mathbb{R}^{6d_h}$. A transformer layer [41] is used for feature fusion and interaction. Eventually, the refined representations make predictions through the fully connected layers.

For the classification task, we employ the standard cross-entropy loss. For the regression task, we use the standard L_1 loss. Combining the task loss $\mathcal{L}_{task}$, separation loss $\mathcal{L}_{sep}$ and adversarial loss $\mathcal{L}_{agn}, \mathcal{L}_{spe}$, the total loss is expressed as follows:

$$\mathcal{L}_{all} = \mathcal{L}_{task} + \alpha \mathcal{L}_{sep} + \beta(\mathcal{L}_{agn} + \mathcal{L}_{spe}), \tag{16}$$

where α and β are the trade-off parameters. Furthermore, we add a gradient reversal layer [9] between the agnostic encoder and the modality discriminator to achieve local optimization of the agnostic adversarial loss $\mathcal{L}_{agn}$.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

We conduct experiments on three standard datasets of multimodal video understanding, including CMU-MOSI [48], CMU-MOSEI [47], and IEMOCAP [3]. These datasets focus on human multimodal emotion recognition and provide unaligned multimodal sequences for each sample. The common protocol of the previous state-of-the-art (SOTA) works [26, 30, 38] is adopted in our experiments.

CMU-MOSI & MOSEI. CMU-MOSI [48] is a dataset containing 2,199 short monologue video clips. Its predetermined data partition has 1,284 samples in the training set, 229 in the validation set, and 686 in the testing set. The acoustic and visual features are extracted at a sampling rate of 12.5 and 15 Hz, respectively. CMU-MOSEI

Table 1: Comparison on the CMU-MOSI dataset.

Approach	$Acc_7 \uparrow$	$Acc_2 \uparrow$	$F1 \uparrow$	$MAE \downarrow$	$Corr \uparrow$
EF-LSTM	31.0	73.6	74.5	1.078	0.542
LF-LSTM	33.7	77.6	77.8	0.988	0.624
RAVEN [44]	31.7	72.7	73.1	1.076	0.544
MCTN [36]	32.7	75.9	76.4	0.991	0.613
MulT [38]	39.1	81.1	81.0	0.889	0.686
PMR [30]	40.6	82.4	82.1	-	-
MICA [26]	40.8	82.6	82.7	-	-
MFSA (ours)	**41.4**	**83.3**	**83.7**	**0.856**	**0.722**

Table 2: Comparison on the CMU-MOSEI dataset.

Approach	$Acc_7 \uparrow$	$Acc_2 \uparrow$	$F1 \uparrow$	$MAE \downarrow$	$Corr \uparrow$
EF-LSTM	46.3	76.1	75.9	0.680	0.585
LF-LSTM	48.8	77.5	78.2	0.624	0.656
RAVEN [44]	45.5	75.4	75.7	0.664	0.599
MCTN [36]	48.2	79.3	79.7	0.631	0.645
MulT [38]	50.7	81.6	81.6	0.591	0.694
PMR [30]	51.8	83.1	82.8	-	-
MICA [26]	52.4	83.7	83.3	-	-
MFSA (ours)	**53.2**	**83.8**	**83.6**	**0.574**	**0.724**

Table 3: Comparison on the IEMOCAP dataset.

Category	Happy		Sad		Angry		Neutral	
Approach	$Acc \uparrow$	$F1 \uparrow$	$Acc \uparrow$	$F1 \uparrow$	$Acc \uparrow$	$F1 \uparrow$	$Acc \uparrow$	$F1 \uparrow$
EF-LSTM	76.2	75.7	70.2	70.5	72.7	67.1	58.1	57.4
LF-LSTM	72.5	71.8	72.9	70.4	68.6	67.9	59.6	56.2
RAVEN [44]	77.0	76.8	67.6	65.6	65.0	64.1	62.0	59.5
MCTN [36]	80.5	77.5	72.0	71.7	64.9	65.6	49.4	49.3
MulT [38]	84.8	81.9	77.7	74.1	73.9	70.2	62.5	59.7
PMR [30]	86.4	83.3	78.5	75.3	75.0	71.3	63.7	60.9
MICA [26]	86.8	83.9	79.3	75.2	75.7	72.4	63.7	61.6
MFSA (ours)	**87.2**	**84.3**	**80.7**	**76.8**	**76.5**	**73.2**	**64.4**	**62.5**

[47] is a dataset made up of 22,856 samples of movie review video clips. Its predetermined data partition has 16,326 samples in the training set, 1,871 in the validation set, and 4,659 in the testing set. The acoustic and visual features are extracted at a sampling rate of 20 and 15 Hz, respectively. For CMU-MOSI & MOSEI, each sample is labeled by human annotators with a sentiment score from -3 of strongly negative to 3 of strongly positive. As in the previous works [26, 30], we adopt diverse metrics including: 7-class accuracy of emotion score classification (Acc_7), binary accuracy of positive/negative emotions (Acc_2), $F1$ score, mean absolute error (MAE), and the pearson correlation ($Corr$).

IEMOCAP . IEMOCAP [3] consists of language, acoustic, and visual modalities from 10 actors recorded in the form of conversations using a Motion Capture camera. Specifically, the multimodal streams consider fixed sampling rate on acoustic (12.5 Hz) and visual (15 Hz) signals. The label annotations consists of four emotions: angry, happy, neutral, and sad. As suggested by [44], 4 emotions (*i.e.*, happy, sad, angry and neutral) are selected for emotion recognition. Following the previous works [39, 44], the classification accuracy (Acc) and $F1$ score are used as evaluation metrics.

Table 4: Results of ablation studies on the CMU-MOSI dataset. "WAL" means Weighted Attention Layer.

Model	$Acc_7 \uparrow$	$Acc_2 \uparrow$	$F1 \uparrow$	$MAE \downarrow$	$Corr \uparrow$
MFSA (Full)	**41.4**	**83.3**	**83.7**	**0.856**	**0.722**
Analysis of Regularization					
w/o $\mathcal{L}_{sep}$	40.9	82.6	83.5	0.864	0.718
w/o $\mathcal{L}_{agn} + \mathcal{L}_{spe}$	39.2	82.4	82.7	0.871	0.711
Importance of Representations					
w/o Modality-Specific	38.6	81.2	81.5	0.898	0.708
w/o Modality-Agnostic	39.2	81.8	82.4	0.881	0.712
Importance of Modules and Strategies					
w/o PSA Module	37.9	80.7	81.3	0.917	0.698
w/o Prediction Chain	40.3	82.5	83.0	0.864	0.718
w/o WAL	40.5	82.7	83.2	0.858	0.720
w/o HCA Module	38.4	81.3	81.8	0.890	0.711
w/o MRU (Mixed-grained)	40.5	82.8	83.2	0.859	0.717
w/o MRU (Coarse-grained)	41.0	83.1	83.4	0.857	0.720
w/o MRU (Fine-grained)	40.2	82.7	83.2	0.862	0.715

4.2 Implementation Details

For the language modality, we convert the transcripts of video into pre-trained Glove word embedding [35] with a 300-dimensional vector. For the acoustic modality, we use COVAREP toolkit [5] for extracting 74-dimensional low-level acoustic features. The features include 12 Mel-frequency cepstral coefficients (MFCCs), voiced segmenting features, glottal source parameters, etc. For the visual modality, the Facet [19] is utilized to indicate 35 facial action units, which records facial muscle movement for representing emotions.

All models are built on the Pytorch toolbox [34] with two Quadro RTX 8000 GPUs. The Adam optimizer [22] is adopted for network optimization. For the CMU-MOSI, MOSEI, and IEMOCAP datasets, the training setting follows: the batch sizes are $\{64, 64, 32\}$, the epochs are $\{100, 100, 60\}$, the learning rates are $\{1e^{-3}, 3e^{-3}, 1e^{-3}\}$, the attention heads are $\{8, 10, 8\}$, the coefficients μ are $\{0.25, 0.15, 0.2\}$, the trade-off parameters α and β are $\{1e^{-2}, 3e^{-2}, 1e^{-2}\}$ and $\{3e^{-2}, 5e^{-2}, 2e^{-2}\}$, respectively. The hidden dimension d is 40 and the output dimension d_h is 128. The number of layers in the PSA module and the HCA module is $M = 5$ and $N = 2$. The hyper-parameters are determined via the validation set.

5 RESULTS AND DISCUSSIONS

5.1 Comparison with State-of-the-Art Methods

We compare the proposed MFSA with recent SOTA works that directly deal with asynchronous multimodal sequences, including Late Fusion LSTM (LF-LSTM), Multimodal Transformer (MulT) [38], Progressive Modality Reinforcement (PMR) [30], and Modality-Invariant Crossmodal Attention (MICA)[26]. To compare the models comprehensively, we employ the Connectionist Temporal Classification (CTC) loss [11] to the prominent approaches (*e.g.*, Early Fusion LSTM (EF-LSTM), Recurrent Attended Variation Embedding Network (RAVEN) [44], and Multimodal Cyclic Translation Network (MCTN) [36]) that cannot be applied directly to the asynchronous multimodal sequence fusion. Concretely, these models train to optimize the CTC alignment objective and the multimodal objective simultaneously.

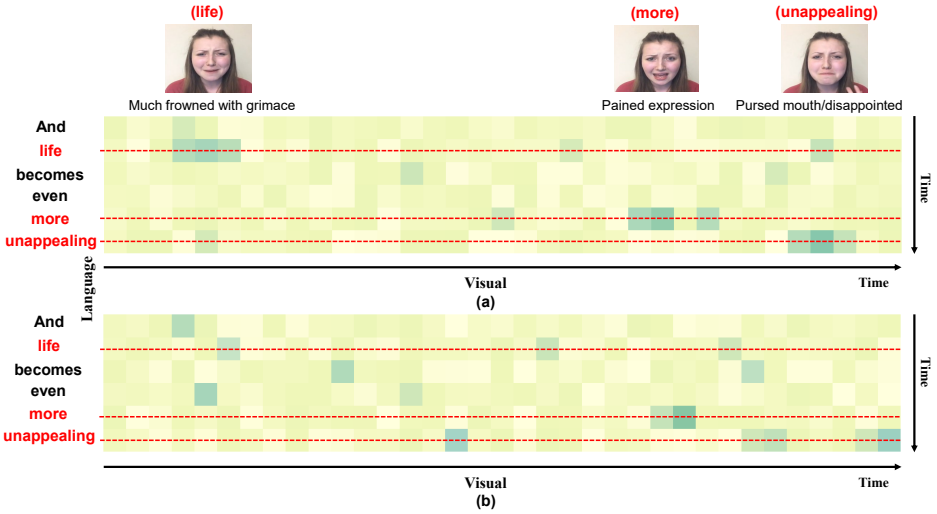
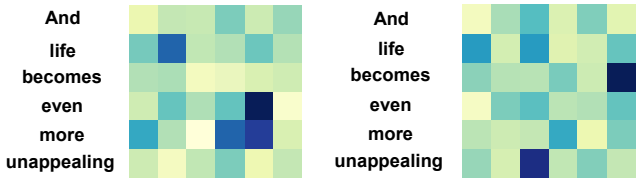


Figure 3: Visualization of the cross-modal attention matrix activation for the proposed HCA module (a) and the SOTA method MulT [38] (b) on the CMU-MOSEI dataset. The textual words which are closely related to emotion are displayed in red. The textual words above the video frames are the corresponding spoken words. Compared to the MulT, our model clearly captures reliable correlations between elements of different modalities. For example, stronger attention is given to the intersection of spoken words that tend to suggest emotions ("unappealing") and facial expression changes in the video ("pursed mouth").



(a) Vanilla self-attention **(b) PSA module**

Figure 4: Visualization of the attention matrix activation from vanilla self-attention [41] and our PSA module in the language sequence.

The experimental results on the three datasets are reported in Tables 1, 2, and 3, respectively. We have the following observations. The proposed MFSA significantly outperforms the existing SOTA approaches [26, 30, 38] without explicit data alignment on all metrics for the three datasets. Meanwhile, our approach provides 8%-12% improvement in most attributes over works [36, 44] that requires CTC. From the perspective of representation learning, not only does our approach surpass works [30, 36, 38, 44] that do not learn distinct representations, but also it is superior to the study [26] that learns cross-modal element correlations only in the modality-invariant space. The above observations demonstrate that it is beneficial to consider both modality-specific and modality-agnostic representations in multimodal learning, as in our MFSA.

5.2 Ablation Studies

We perform thorough ablation studies on the CMU-MOSI dataset to verify the necessity of the proposed components. Experimental results in Table 4 display the following observations.

Analysis of Regularization . Regularization plays a critical role in learning the distinct representations. To quantitatively verify the importance of the proposed regularization, we remove either loss separately to perform the experiments. When the separation loss $\mathcal{L}_{sep}$ is removed, the decreased results suggest that it is beneficial to perform orthogonal constraint between the different representations. When the adversarial losses ($\mathcal{L}_{agn}$ + $\mathcal{L}_{spe}$) are removed, the training process of the model does not involve the double-discriminator adversarial strategy. In this case, the poor performance clearly demonstrates the advantage of the adversarial manner in learning distinct multimodal representations.

Importance of Representations . We observe the model's performance by using only either representation in the feature fusion phase. The poor performance of both shows the effectiveness of learning modality-specific and -agnostic representations across multiple modalities. Furthermore, the worse performance without the modality-specific representations inspire us to depict multimodal representations from the perspective of feature disentanglement and focus on the unique characteristics of each modality.

Importance of Modules and Strategies. Finally, we evaluate the effectiveness of the different modules and strategies. Firstly, when the predictive self-attention (PSA) module and the hierarchical cross-modal attention (HCA) module are removed, there is a significant drop on the performance of the model. These observations suggest that reinforcing multimodal representations via self-attention and cross-modal attention mechanisms is indispensable. Meanwhile, we find that both the Weighted Attention Layer (WAL) and the convolution-based prediction chain provide significant contributions to improve performance. For the HCA module,

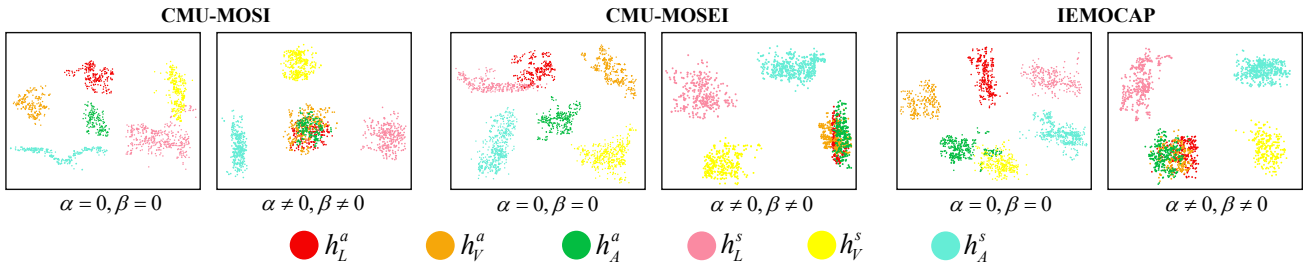


Figure 5: Visualization of the modality-specific and modality-agnostic representations in the testing set on three benchmarks. $\alpha = 0, \beta = 0$ denotes without separation and adversarial losses, and vice versa. The light colors correspond to specific parts, while dark colors correspond to agnostic parts.

the mixed-grained, coarse-grained, and fine-grained modality reinforcement units (MRU) respectively are removed to explore the gains from the hierarchical structure. The experimental results in the bottom half of Table 4 show that it is beneficial to consider cross-modal interactions at different granularities. Furthermore, we observe that the models are sensitive to fine-grained and mixed-grained MRU units. The results inspire us to focus more on cross-modal interactions between independent modalities and integrated interactions across multiple modalities.

5.3 Visualization and Analysis

Effectiveness of the HCA module. We visualize an example on the CMU-MOSEI dataset to understand how the proposed HCA module works when modelling cross-modal interactions. Figure 3.(a) & (b) show the attention matrix activation for the last layer of the fine-grained MRU in our HCA module and the last layer in the SOTA method MulT [38] (Only MulT is open source), respectively. Since deeper parts indicate stronger attention, we find that our module learns a reasonable correlation between the video frames and the spoken words. The emotion-related words (*e.g.*, "life", "unappealing") successfully attend to the video frames that contain the corresponding facial expression changes (*e.g.*, "much frowned with grimace" and "pursed mouth"). Compared to the MulT, our approach can encourage the model to focus on more meaningful signals and elements across two modalities. A reasonable explanation is that the branch of cross-modal interactions is guided and supervised by parameter updates when learning the modality-agnostic representations, resulting in latent distribution alignment across modalities. It is known that mitigating distribution discrepancy can effectively improve cross-modal correlations [26].

Effectiveness of the PSA module. To prove the superiority of our model in learning contextual dependencies, we visualize the attention matrix activation in the last layer using the self-attention [41] and the proposed PSA module, respectively. As shown in Figure 4.(a), the self-attention only focuses on the phrase "even more", leading to a meaningless correlation. In contrast, our PSA module attends to the relationship between "becomes" and "unappealing", which correctly captures the modality-specific context semantics (*i.e.*, linking verb + predicative) of the language modality in Figure 4.(b). These observations suggest that the attention pattern incorporating convolutional inductive bias favours complementary relationships that emphasize cross-modal correlations with those in the HCA module. The phenomenon may benefit from projection constraints when learning the modality-specific representations.

Visualization of Distinct Representations. Understanding the distributions of distinct representations plays an essential role in feature disentanglement. In Figure 5, we visualize the modality-specific representations $\boldsymbol{h}_m^s$ and modality-agnostic representations $\boldsymbol{h}_m^a$ learned in the testing samples of the three datasets, where $m \in \{L, V, A\}$. When there is no regularization constraints (*i.e.*, $\alpha = 0, \beta = 0$), the modality-agnostic representations are not learned, and the distributions of $\boldsymbol{h}_m^s$ and $\boldsymbol{h}_m^a$ are occasionally blurred. Conversely, when $\alpha \neq 0, \beta \neq 0$, the distributions of $\boldsymbol{h}_m^a$ are mixed together, where adversarial training effectively aligns distributions of different modalities and minimizes the modality gap. In addition, the modality-specific representations of different modalities are well separated, and their distributions become more compact. These observations clearly demonstrate that our approach can capture both the commonality and diversity across multiple modalities.

6 CONCLUSION

In this paper, we propose a Multimodal Fusion approach for learning modality-Specific and modality-Agnostic representations (MFSA) to refine multimodal representations and leverage the complementarity among different modalities. On the one hand, our MFSA fully explores the unique characteristics and diversity of each modality over the modality-specific spaces. On the other hand, the proposed approach effectively mitigates the modality gap and captures the commonality across modalities over the modality-agnostic space. These distinct representations provide new insight and perspective for performing effective feature fusion and interaction in asynchronous multimodal sequences. Numerous experiments prove the effectiveness of the proposed modules. Furthermore, the insight of feature disentanglement can be easily extended to other tasks.

ACKNOWLEDGMENTS

This work is supported by National Key R&D Program of China (2021ZD0113502), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0103) and National Natural Science Foundation of China under Grant (82090052).

REFERENCES

[1] Hassan Akbari, Linagzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv preprint arXiv:2104.11178* (2021).

[2] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. *Advances in neural information processing systems* 29 (2016).

[3] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335–359.

[4] Zhaoyu Chen, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Wenqiang Zhang. 2022. Towards Practical Certifiable Patch Defense with Vision Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15148–15158.

[5] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 960–964.

[6] Hazarika Devamanyu, Zimmermann Roger, and Poria Soujanya. 2020. MISA: Modality-invariant and-specific representations for multimodal sentiment analysis.. In *Proceedings of the 28th ACM International Conference on Multimedia*, Vol. 34. 1122–1131.

[7] Wenting Duan, Lei Zhang, Jordan Colman, Giosue Gulli, and Xujiong Ye. 2021. Multi-modal Brain Segmentation Using Hyper-Fused Convolutional Neural Network. In *International Workshop on Machine Learning in Clinical Neuroimaging*. Springer, 82–91.

[8] Quan Gan, Shangfei Wang, Longfei Hao, and Qiang Ji. 2017. A multimodal deep regression bayesian network for affective video content analyses. In *Proceedings of the IEEE International Conference on Computer Vision*. 5113–5122.

[9] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*. PMLR, 1180–1189.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[11] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*. 369–376.

[12] P. Hai, T. Manzini, P. P. Liang, and Barnabás Poczós. 2018. Seq2Seq2Sentiment: Multimodal Sequence to Sequence Models for Sentiment Analysis. (2018).

[13] Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412* (2021).

[14] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).

[15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[16] J. Hu, J. Lu, and Y. P. Tan. 2018. Sharable and Individual Multi-View Metric Learning. *Pattern Analysis & Machine Intelligence IEEE Transactions on* 40, 9 (2018), 2281–2288.

[17] Hao Huang, Yongtao Wang, Zhaoyu Chen, Yuze Zhang, Yuheng Li, Zhi Tang, Wei Chu, Jingdong Chen, Weisi Lin, and Kai-Kuang Ma. 2022. CMUA-Watermark: A Cross-Model Universal Adversarial Watermark for Combating Deepfakes. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 1 (Jun. 2022), 989–997. https://doi.org/10.1609/aaai.v36i1.19982

[18] Yibo Huang, Hongqian Wen, Linbo Qing, Rulong Jin, and Leiming Xiao. 2021. Emotion Recognition Based on Body and Context Fusion in the Wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3609–3617.

[19] iMotions. 2017. *Facial expression analysis*.

[20] Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186* (2019).

[21] Phil Kim. 2017. Convolutional neural network. In *MATLAB deep learning*. Springer, 121–147.

[22] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[23] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. 2019. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10143–10152.

[24] Xiang Li, Chao Wang, Jiwei Tan, Xiaoyi Zeng, Dan Ou, Dan Ou, and Bo Zheng. 2020. Adversarial multimodal representation learning for click-through rate prediction. In *Proceedings of The Web Conference 2020*. 827–836.

[25] Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. *arXiv preprint arXiv:1808.03920* (2018).

[26] Tao Liang, Guosheng Lin, Lei Feng, Yan Zhang, and Fengmao Lv. 2021. Attention Is Not Enough: Mitigating the Distribution Discrepancy in Asynchronous Multimodal Sequence Fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8148–8156.

[27] Siao Liu, Zhaoyu Chen, Wei Li, Jiwei Zhu, Jiafeng Wang, Wenqiang Zhang, and Zhongxue Gan. 2022. Efficient Universal Shuffle Attack for Visual Object Tracking. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2739–2743. https://doi.org/10.1109/ICASSP43922.2022.9747773

[28] Yang Liu, Jing Liu, Mengyang Zhao, Shuang Li, and Liang Song. 2022. Collaborative Normality Learning Framework for Weakly Supervised Video Anomaly Detection. *IEEE Transactions on Circuits and Systems II: Express Briefs* 69, 5 (2022), 2508–2512. https://doi.org/10.1109/TCSII.2022.3161061

[29] Yang Liu, Jing Liu, Xiaoguang Zhu, Donglai Wei, Xiaohong Huang, and Liang Song. 2022. Learning Task-Specific Representation for Video Anomaly Detection with Spatial-Temporal Attention. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2190–2194. https://doi.org/10.1109/ICASSP43922.2022.9746822

[30] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021. Progressive Modality Reinforcement for Human Multimodal Emotion Recognition From Unaligned Multimodal Sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2554–2562.

[31] Dung Nguyen, Kien Nguyen, Sridha Sridharan, David Dean, and Clinton Fookes. 2018. Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition. *Computer Vision and Image Understanding* 174 (2018), 33–42.

[32] Dung Nguyen, Kien Nguyen, Sridha Sridharan, Afsane Ghasemi, David Dean, and Clinton Fookes. 2017. Deep spatio-temporal features for multimodal emotion recognition. In *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 1215–1223.

[33] G. Park and W. Im. 2016. Image-Text Multi-Modal Representation Learning by Adversarial Backpropagation. (2016).

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[35] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[36] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6892–6899.

[37] Z. Sun, P. Sarma, W. Sethares, and Y. Liang. 2020. Learning Relationships between Text, Audio, and Video via Deep Canonical Correlation for Multimodal Language Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 5 (2020), 8992–8999.

[38] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2019. NIH Public Access, 6558.

[39] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176* (2018).

[40] Yao-Hung Hubert Tsai, Martin Q Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Multimodal routing: Improving local and global interpretability of multimodal language analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, Vol. 2020. NIH Public Access, 1823.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[42] Shunli Wang, Dingkang Yang, Peng Zhai, Chixiao Chen, and Lihua Zhang. 2021. Tsa-net: Tube self-attention network for action quality assessment. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4902–4910.

[43] Weiran Wang, Xinchen Yan, Honglak Lee, and Karen Livescu. 2016. Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454* (2016).

[44] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7216–7223.

[45] Yujing Wang, Yaming Yang, Jiangang Bai, Mingliang Zhang, Jing Bai, Jing Yu, Ce Zhang, and Yunhai Tong. 2020. Predictive Attention Transformer: Improving Transformer with Attention Map Prediction. (2020).

[46] Yang Wu, Zijie Lin, Yanyan Zhao, Bing Qin, and Li-Nan Zhu. 2021. A Text-Centered Shared-Private Framework via Cross-Modal Prediction for Multimodal Sentiment Analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 4730–4738.

[47] Amir Zadeh and Paul Pu. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*.

[48] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems* 31, 6 (2016), 82–88.

[49] Zhihong Zeng, Jilin Tu, Brian Pianfetti, Ming Liu, Tong Zhang, Zhenqiu Zhang, Thomas S Huang, and Stephen Levinson. 2005. Audio-visual affect recognition through multi-stream fused HMM for HCI. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 2. IEEE, 967–972.

[50] Peng Zhai, Jie Luo, Zhiyan Dong, Lihua Zhang, Shunli Wang, and Dingkang Yang. 2022. Robust Adversarial Reinforcement Learning with Dissipation Inequation Constraint. (2022).

[51] S. F. Zhang, J. H. Zhai, B. J. Xie, Y. Zhan, and X. Wang. 2019. Multimodal Representation Learning: Advances, Trends and Challenges. In *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*.

[52] Y. Zhang, M. Chen, J. Shen, and C. Wang. 2022. Tailor Versatile Multi-modal Learning for Multi-label Emotion Recognition. *arXiv e-prints* (2022).

[53] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. 2016. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3438–3446.