Scale-Selectable Global Information and Discrepancy Learning Network for Multimodal Sentiment Analysis

Xiaojiang He, Yushan Pan *Senior Member, IEEE*, Xinfei Guo *Senior Member, IEEE*, Zhijie Xu *Senior Member, IEEE*, Chenguang Yang *Fellow, IEEE*

***Abstract*—Multimodal sentiment analysis and depression detection are pivotal for advancing human-computer interaction, yet significant challenges remain. First, the limited extraction of global contextual information within individual modalities risks the loss of modal-specific features. Second, existing methods often prioritize unaligned textual interactions, neglecting critical inter-modal discrepancies. To address these issues, we propose the Scale-Selectable Global and Discrepancy Learning Network (SSGDL), an innovative framework that integrates two core modules: the Cross-Shaped Dynamic Scale Attention Module (CS-DSA) and the Primary-Secondary modal Discrepancy Learning Module (PS-MDL). The CS-DSA dynamically selects scales and employs cross-shaped attention to capture comprehensive global context and intricate internal correlations, effectively producing a fused modal representation. Meanwhile, the PS-MDL designates the fused modal as primary and utilizes cross-attention mechanisms to learn discrepancy representations between it and other modalities (textual, acoustic, and visual). By leveraging inter-modal discrepancies, SSGDL achieves a more nuanced and holistic understanding of emotional content. Extensive experiments on three benchmark multimodal sentiment analysis datasets (MOSI, MOSEI, SIMS) and a depression detection dataset (AVEC2019) demonstrate that SSGDL consistently outperforms state-of-the-art approaches, setting a new benchmark for multimodal affective computing.**

***Index Terms*—Multimodal Sentiment Analysis, depression detection, Scale-Selectabl Global Information, Inter-modal Discrepancy Learning, Neuro-scientific theories.**

I. Introduction

Sentiment plays a crucial role in human cognition, particularly in decision-making, perception, and interpersonal communication. It can be inferred from various sources of information, including speech, facial expressions, text, body movements, and physiological signals, with each source representing a distinct modal. Moreover, sentiment analysis serves as a key enabler in bridging the gap between artificial intelligence (AI) and affective computing, allowing machines to more accurately understand and respond to human emotions. Early research efforts predominantly focused on single-modal sentiment analysis, including text-based [1] [2], image-based [3] [4], and audio-based sentiment analysis [5] [6]. While these approaches have demonstrated utility in specific domains, single-modal sentiment analysis models inherently suffer from limitations such as sensitivity to noise, susceptibility to bias, and the generation of ambiguous or contradictory results. To address these shortcomings, Multimodal Sentiment Analysis (MSA) has emerged as a robust alternative by integrating data from multiple modalities, such as text, audio, and visual signals-as illustrated in Fig. 1. By leveraging the complementary strengths of these modalities, MSA provides a more comprehensive and accurate representation of the complexity and diversity inherent in real-world emotional expressions [15] [16] [17].

Recent advancements in multimodal fusion techniques have further propelled the field of MSA. Common fusion strategies include feature-level fusion [7] [8], decision-level fusion [9] [10], and consistent regression fusion [11] [12]. Furthermore, state-of-the-art attention-based models, such as Multi-Attention Recurrent Networks (MARNs) [13], have shown significant progress in capturing both intra-modal and inter-modal dynamics. Beyond these efforts, Wu et al. [14] introduced an innovative mechanism for detecting word-level inconsistency, while Wen et al. [57] proposed a hardware-optimized architecture for long short-term memory networks (MLSTM) based on memristor technology. Li et al. [56] further advanced the field by leveraging raw multimodal data for pre-training, facilitating deeper exploration of multimodal information, enhancing model generalization capabilities, and substantially reducing manual labelling costs. As demonstrated by Qureshi et al. [18], there is a strong correlation between sentiment analysis and video-based depression diagnosis, which enables the application of shared techniques such as temporal modelling with LSTM architectures, multimodal fusion methods (e.g., attention-based and tensor fusion), and feature extraction strategies like Mel-frequency cepstral coefficients (MFCC) for acoustic signals and lexical embedding models for textual inputs. These methods are widely used in both domains to detect nuanced emotional patterns and behavioral cues.

Despite recent advances, traditional multimodal sentiment analysis (MSA) methods still exhibit fundamental limitations.

This work was supported by the Leadership Talent Program(Science and Education) of SIP, KJL2024104, Xi'an Jiaotong-Liverpool University RDF-21-02-008 and Jiangsu Double-Innovation Plan No. JSSCBS20230474. (*Corresponding authors: Yushan Pan*)

X. He is a Research Assistant with the University of Liverpool and Xi'an Jiaotong - Liverpool University, Suzhou, 215123, China (E-mail: hexiaojiang2022@163.com)

Y. Pan is is an Assistant Professor with Xi'an Jiaotong-Liverpool University(E-mail: Yushan.Pan@ieee.org)

X. Guo is an Associate Professor with Shanghai Jiao Tong University(E-mail: xinfei.guo@sjtu.edu.cn)

Z. Xu is a Professor with Xi'an Jiaotong-Liverpool University(E-mail: Zhijie.Xu@xjtlu.edu.cn)

C. Yang is a Professor with the University of Liverpool(E-mail: cyang@ieee.org)

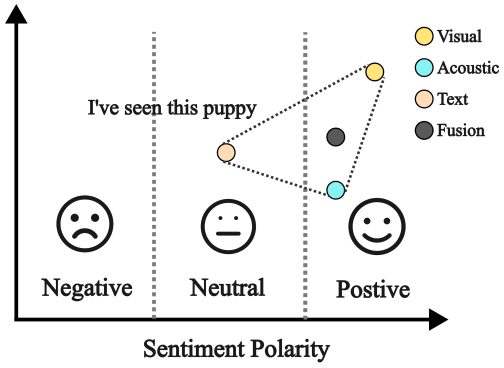


Fig. 1. Multimodal Data Can Provide More Accurate Information for Sentiment Analysis

Notably, those existing approaches rely on fixed-scale feature extraction strategies [32], which lack the capacity to dynamically adapt to evolving cross-modal dependencies. This static modelling strategy often leads to a loss of semantically rich contextual information during fusion, particularly when modalities are weakly aligned or temporally asynchronous. For instance, emotional shifts in vocal tone may precede or lag behind corresponding facial expressions—temporal dynamics that static fusion windows fail to capture effectively. Furthermore, although several studies have explored inter-modal interactions [13] [14], they typically assume modal consistency and overlook the presence of inherent cross-modal contradictions—such as sarcasm, irony, depressive indicators, or affective ambiguity. These subtle emotional expressions (e.g., a cheerful facial expression paired with a sarcastic tone) demand explicit discrepancy modelling. Additionally, methods such as [31] attempt to enhance inter-modal interactions, but often neglect fine-grained discrepancies that are essential for revealing nuanced emotional cues and underlying speaker intentions. Thus, previous approaches either ignore these complex signals or treat them as noise, resulting in coarse-grained sentiment representations that fail to capture the intricacy and variability of real-world emotional expression.

To overcome these challenges, we propose the Scale-Selectable Global Information and Discrepancy Learning Network (SSGDL), which introduces two novel modules: the Cross-Shaped Dynamic Scale Attention Module (CS-DSA) and the Primary-Secondary Modality Discrepancy Learning Module (PS-MDL). The CS-DSA dynamically adjusts the receptive field scale based on the input stimulus, employing a cross-shaped attention mechanism to capture contextual information at varying scales. This enables the model to effectively aggregate global information and modal-specific correlations, thereby generating robust fused modal representations. The PS-MDL further enhances the fused modal representation by treating it as the primary modal and designating the other modalities (e.g., text, acoustic, visual) as auxiliary. It leverages cross-attention and self-attention mechanisms to hierarchically integrate the auxiliary modalities, effectively extracting critical discrepancy information that highlights the complementarity and uniqueness of each modal.

By combining these innovations, our approach not only addresses the limitations of traditional fusion strategies but also provides a more nuanced and holistic understanding of emotional content. Extensive experiments on benchmark datasets (MOSI, MOSEI, SIMS, and AVEC2019) demonstrate that SSGDL consistently outperforms state-of-the-art methods, setting a new benchmark for multimodal affective computing.

Thus, our contributions include:

- We contribute an innovative framework, the Scale-Selectable Global and Differential Learning Network (SSGDL). By effectively leveraging cross-modal differences, SSGDL enables a more nuanced and comprehensive understanding of emotional content. This approach enhances the model's ability to capture intricate interdependencies between modalities, fostering a deeper integration of multimodal information.
- We introduce the Cross-Shaped Dynamic Scale Attention (CS-DSA) module. This efficient attention mechanism automatically selects the most appropriate kernel size and employs cross-shaped interactions to comprehensively extract essential information from each neuron while assessing correlations, thereby uncovering complex interrelationships in multimodal data.
- We design a Primary-Secondary Modality Discrepancy Learning Module (PS-MDL) to capture discordant information. This network structure includes primary and auxiliary modal generation, cross-attention, and self-attention mechanisms specifically designed to capture and utilize inter-modal discrepancy information.
- Extensive experiments are conducted on three multimodal sentiment analysis datasets—SIMS, CMU-MOSI, and CMU-MOSEI—to thoroughly evaluate the superiority and effectiveness of the proposed method. Furthermore, to assess the robustness of the model across different domains, additional experiments are performed on the cross-domain AVEC 2019 depression detection dataset, yielding exceptional results.

II. Related Work

A. Multimodal Sentiment Analysis

Initial sentiment analysis research was heavily centred on textual data to assess users' emotional orientation (positive, negative, or neutral) [19] [20]. MSA broadens this traditional approach by integrating speech and visual features to more comprehensively capture the sentiment expressed in an utterance. Research in MSA mainly focuses on two areas: representation learning and multimodal fusion. In the domain of unimodal representation learning, Sun et al. [21] use utterance-level representations from each modal as a global multimodal context, which interacts with local unimodal features for mutual enhancement. Cristina et al. [22] introduced CrowdDM, a sentiment analysis-guided group decision-making model that utilizes crowd intelligence from social networks to solve decision-making challenges.

Regarding multimodal fusion, three main strategies have been explored. We note that the categorization of fusion

strategies in the previous section—namely feature-level fusion, decision-level fusion, and consistent regression fusion—originates from a signal-processing perspective and focuses on the level of fusion. In this section, we adopt a complementary perspective that emphasizes when and how fusion occurs during model training. Accordingly, the three strategies discussed below—early, late, and hybrid fusion—can be seen as architectural counterparts to the earlier categorization. Specifically, early fusion typically corresponds to feature-level fusion; late fusion aligns with decision-level fusion; hybrid fusion often incorporates ideas from consistent regression or multi-stage supervision.

Early fusion combines features from different modalities at the initial stages, allowing the model to learn joint representations from the start and enabling interaction between modalities. For example, Tsai et al. [23] introduced a cross-modal Transformer that enhances the target modal through cross-modal attention, generating unified representations early in the learning process. Late fusion, on the other hand, integrates modalities after they have been processed independently, typically through methods like concatenation or tensor-based approaches. Zadeh et al. [24] developed a tensor fusion network that computes the outer product of unimodal representations, capturing cross-modal interactions at a later stage. Additionally, hybrid approaches combine the strengths of both early and late fusion. Li et al. [25] proposed a hierarchical disentanglement technique that effectively extracts shared and private sentiment information from different modalities. Meanwhile, Wu et al. [58] proposed the Multimodal Multi-loss Fusion Network (MMML), which integrates audio and text signals using a transformer-based fusion network. While MMML focuses on optimizing the fusion process through multi-task learning, our approach introduces a novel Cross-Shaped Dynamic Scale Attention (CS-DSA) module, which dynamically adjusts the receptive field scale to capture both global and local contextual information. This dynamic scale selection mechanism distinguishes our work from MMML and allows for more flexible and adaptive feature extraction.

Despite recent advancements, existing studies often overlook the heterogeneity across modalities—that is, the naturally different ways in which text, audio, and visual cues express sentiment [33]. This heterogeneity is critical for understanding emotionally ambiguous or contradictory scenarios, such as a sarcastic remark with a smiling face but a flat tone. Moreover, current datasets often contain spurious correlations between sentiment labels and superficial features (e.g., certain facial expressions co-occurring with positive labels regardless of actual sentiment), which can mislead models into learning dataset-specific biases rather than genuine emotional understanding. These challenges increase the risk of intra-modal redundancy, inter-modal information loss, and incorrect generalization. Therefore, it is essential to design models that can both respect modal-specific characteristics and explicitly account for misleading cross-modal correlations. Our proposed discrepancy-aware fusion module (PS-MDL) directly addresses this need by modelling primary-secondary differences among modalities to mitigate such issues.

B. Transformer-Based Multimodal Interaction

Transformer-based networks have shown promise in modelling global contextual information, but important distinctions are often overlooked. Specifically, intra-modal global context refers to long-range dependencies within individual modalities, such as emotional flow in textual discourse or tonal variation in acoustic signals [26] [28]. Meanwhile, inter-modal global context involves capturing coherent, aligned semantics across modalities, which is essential for resolving cross-modal ambiguities and enhancing emotional understanding [27] [30].

However, those Transformer-based multimodal approaches still suffer from two key limitations. First, they often apply the same fusion mechanism across modalities—a problem emphasized by [34] —which ignores the heterogeneity and temporal misalignment inherent in real-world multimodal data. Second, although models like the Multimodal Transformer [27] and tensor-based methods [30] use cross-modal attention, they fail to adaptively adjust to modality-specific dynamics or distinguish between reliable and misleading cross-modal cues. This can lead to spurious correlations, as noted in [29], where models learn dataset-specific patterns (e.g., facial expressions statistically tied to labels) instead of genuine emotional semantics.

These issues motivate the need for a model that can selectively aggregate both intra-modal and inter-modal global information, while accounting for modality discrepancies. Our CS-DSA module is designed to address this gap by dynamically adjusting attention scales to model contextual relationships along both temporal and modal dimensions.

III. Methodology

In this section, we will provide a detailed explanation of the proposed model and its intricate architectural components. Both multimodal sentiment analysis rely on extracting relevant information from different modalities to assess potential inconsistencies.

A. Task Setup

The acoustic (a), visual (v), and textual (t) modalities from the same video segments are used for determining sentiment polarity, and the fusion modal (f) generated by before three modalities. These modalities can be represented as $I_m \in \mathbb{R}^{T_m \times d_m}$, where T_m denotes the sequence length, d_m represents the dimensionality of each modal, and $m \in \{a, v, t, f\}$.

B. Overall Architecture

As shown in Fig. 2, the Scale-Selectable Global Information and Discrepancy Learning Network (SSGDL) for multimodal sentiment analysis includes four main components: feature extraction, the Cross-Shaped Dynamic Scale Attention Module (CS-DSA), the Primary-Secondary Modality Discrepancy Learning Module (PS-MDL), and the prediction layer. The CS-DSA layer employs a novel attention mechanism to explore complex intra- and inter-modal relationships, generating a new fused modal. The PS-MDL layer then captures discrepancies

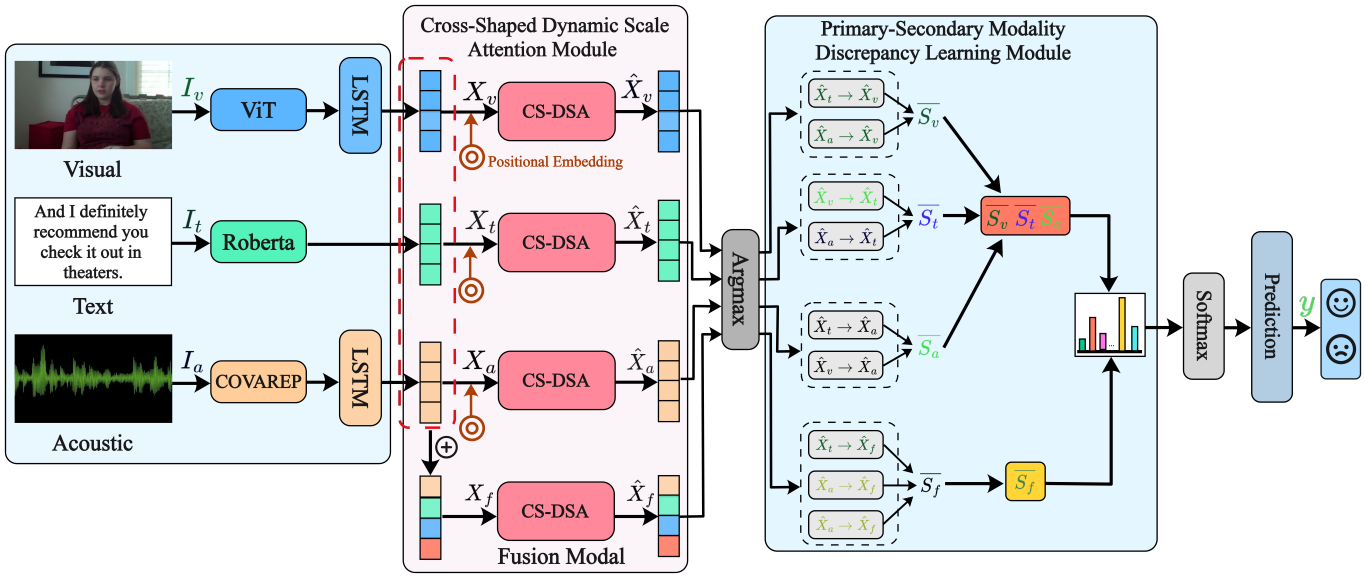


Fig. 2. The proposed SSGDL model framework.

between the primary and auxiliary modalities using cross-attention, which is further refined with self-attention. Finally, a sequence model in the prediction layer forecasts sentiment outcomes based on all modal pairs.

C. Feature Extraction

1) Word Embedding: For the verbal modal, we selected a modified and optimized version of BERT, specifically the 12-layer RoBERTa, as our text encoder. The text is initially processed through RoBERTa's tokenizer, which adds two special tokens: [CLS] at the beginning and [SEP] at the end of the sentence. These tokens help the model identify sentence boundaries and assign contextually relevant representations to each word. The resulting segmentation sequence $I_m \in \mathbb{R}^{T_m \times d_m}$ from the tokenizer is then utilized for further processing in RoBERTa.

2) Visual Feature: For the visual modal, we utilize a pre-trained Vision Transformer (ViT) as our visual encoder, focusing on facial expressions as the primary medium of sentiment conveyance. Since certain facial features, particularly the eyes and mouth, more precisely reflect emotional states, the ViT is leveraged to extract both global and localized facial details. This method allows for a thorough capture of sentiment-related signals from facial expressions. The visual modal representation is expressed as X_v.

3) Acoustic Feature: For the acoustic modal, we utilize the COVERAP analysis framework to extract a range of handcrafted acoustic features. These features encompass 12 Mel-frequency cepstral coefficients, pitch, volume, glottal source parameters, and additional vocal attributes pertinent to emotional expression and tone. By employing the CMU-MultimodalSDK, we generate COVERAP feature sequences for each sample within the multimodal dataset, facilitating an in-depth examination of auditory data. In this study, the acoustic modal representation is denoted as X_a.

D. Cross-Shaped Dynamic Scale Attention Module

While the MMML model employs a fixed cross-modal attention mechanism, our CS-DSA module introduces a dynamic scale selection mechanism that adaptively adjusts the receptive field based on the input stimulus. This allows our model to capture both global and local contextual information more effectively. Additionally, the cross-shaped attention mechanism in CS-DSA enables neurons to acquire dense contextual information from all other neurons, further enhancing the model's ability to handle long-range dependencies. These innovations distinguish our approach from MMML and contribute to the superior performance of our model on benchmark datasets.

The CS-DSA is designed to extract comprehensive global contextual information while capturing complex intra-modal and inter-modal relationships. Unlike traditional methods that rely on fixed-scale feature extraction, the CS-DSA dynamically adjusts the receptive field scale based on the stimulus level, inspired by the adaptive modulation of neuronal receptive fields in the visual cortex. This dynamic scale selection is achieved through a novel gating mechanism that integrates information from multiple branches, each containing data at different scales. Additionally, the CS-DSA employs a cross-shaped attention mechanism to capture contextual information along both the horizontal and vertical axes of the feature map, enabling neurons to acquire dense contextual information from all other neurons. This dual-loop cross-shaped attention mechanism significantly enhances the model's ability to capture global and local dependencies, addressing the limitations of traditional attention mechanisms in handling long-range dependencies. Specifically, the CS-DSA is operationalized through four key processes—Divide, Cross-Interact, Fuse, and Select—as depicted in Fig. 3. This figure illustrates a dual-branch setup with kernels of different sizes, though the approach can be extended to include multiple branches.

1) Divide: For any feature map $X_m \in \mathbb{R}^{C' \times W' \times H'}$, we first apply two distinct transformations: $\tilde{F} : X \rightarrow \tilde{X} \in$

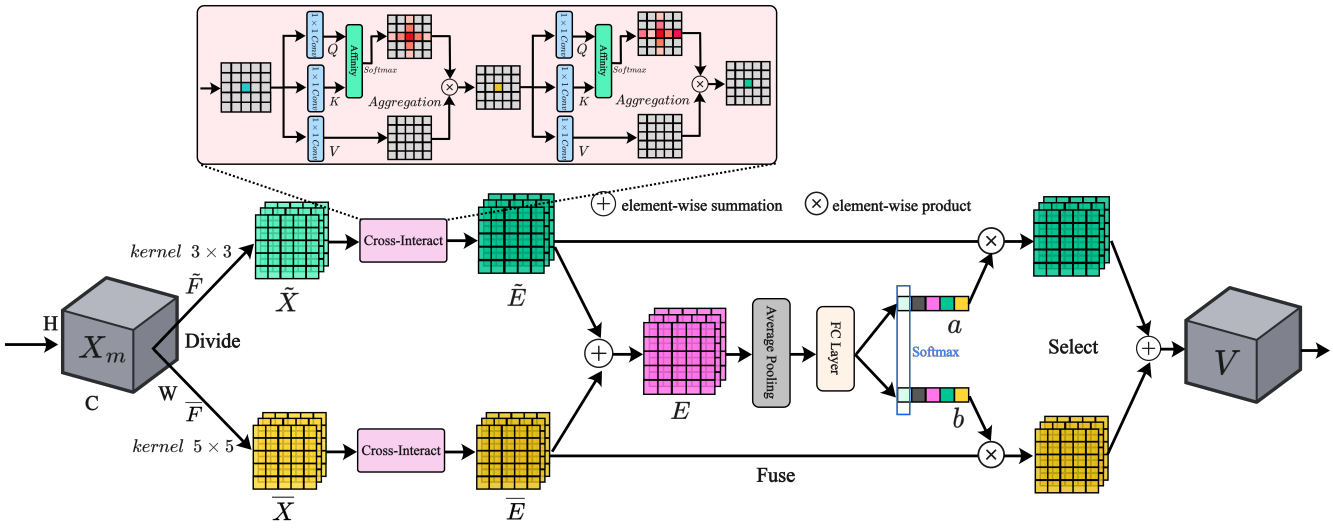


Fig. 3. The proposed CS-DSA model framework.

$\mathbb{R}^{C\times W\times H}$ and $\overline{F}: X \rightarrow \overline{X} \in \mathbb{R}^{C\times W\times H}$, utilizing kernel sizes of 3 and 5, respectively. It is essential to highlight that both $\tilde{F}$ and $\overline{F}$ leverage efficient grouped or depth-wise convolutions, followed by Batch Normalization and the ReLU activation function in sequence. To enhance efficiency, the traditional 5×5 convolution is substituted with a 3×3 dilated convolution, featuring a dilation rate of 2.

2) Interact: The interaction module incorporates an advanced cross-shaped attention mechanism to capture contextual information along both the horizontal and vertical axes of the feature map, enhancing neuron-level representational power. This is achieved through a series of convolution operations and attention mechanisms applied to the feature map.

First Pass: Given a local feature map $\overline{X} \in \mathbb{R}^{C\times W\times H}$, the module first applies two 1×1 convolutions to generate the query Q and key K feature maps, where $\{Q, K\} \in \mathbb{R}^{C'\times W\times H}$. Here, C' represents a reduced channel count, lower than C, enabling effective dimensionality reduction. The attention map $A \in \mathbb{R}^{(H+W-1)\times W\times H}$ is then computed by performing an affinity operation at each spatial location u within feature map Q. This operation compares the feature vector Q_u at position u with a set of feature vectors from K aligned along the same row or column as u, and the affinity score is computed as:

$$d_{i,u} = Q_u \cdot \Omega_{i,u}^{\top} \tag{1}$$

where $\Omega_{i,u}$ represents the feature vectors from K that align with Q_u along the same row or column. The soft-max operation is then applied across the resulting matrix D, producing the attention map A.

Next, a second 1×1 convolution generates the adaptive feature map V, and contextual information is aggregated from neighbouring features in V via the following equation:

$$\overline{X}'_u = \sum_{i\in|\Phi_u|} A_{i,u}\Phi_{i,u} + \overline{X}_u \tag{2}$$

where Φ_u consists of feature vectors from V that are aligned along the same row or column as u, and $A_{i,u}$ is the corresponding attention weight. This process is illustrated in Fig. 3, where the flow of information during the first pass is shown.

Second Pass: To further enhance the contextual information, a second pass of cross-shaped attention is applied. This pass takes the feature map $\overline{X}'$ from the first pass as input and generates the updated feature map $\overline{X}''$, as shown in Fig. 4.

The attention map for the second pass, denoted A', is computed similarly to the first pass, but now applied to the updated feature map $\overline{X}'$. This second pass allows the model to propagate contextual information across all spatial dimensions, capturing more global and non-local contextual relationships.

For any position u in $\overline{X}''$, if u and another position θ are aligned in the same row or column, the update rule is:

$$\overline{X}''_u \leftarrow f(A', u, \theta) \cdot \overline{X}_\theta \cdot [f(A, u, \theta) + 1] \tag{3}$$

where $f(A, u, \theta)$ represents the function mapping attention weights to the spatial relationship between u and θ. When u and θ are not aligned in the same row or column, the update rule becomes:

$$\begin{aligned}\overline{X}''_u \leftarrow [f(A', u_x, u_y, \theta_x, u_y) \cdot f(A, \theta_x, u_y, \theta_x, \theta_y) + \\ f(A', u_x, u_y, \theta_x, \theta_y) \cdot f(A, u_x, \theta_y, \theta_x, \theta_y)] \cdot \overline{X}_\theta\end{aligned} \tag{4}$$

This second pass enables the model to capture richer and more global contextual information, overcoming the limitations of the first pass by addressing both aligned and non-aligned neurons. Additionally, for a more concise representation, we provide the algorithmic flow of Interact in Algorithm 1.

3) Fusion: As outlined earlier, our aim is to provide neurons with the ability to dynamically adjust their receptive field (RF) sizes in response to the content of the input stimulus. To accomplish this, we employ a gating mechanism that controls the flow of information from various branches, each containing data at different scales, into the neurons in the subsequent layer. The gate must be capable of integrating information from all these branches. Initially, the outputs from multiple

Algorithm 1 The Algorithm Flow of Interact in the CS-DSA Module

Input: Feature map $\overline{X} \in \mathbb{R}^{C\times W\times H}$
Output: Enhanced feature map $\overline{X}'' \in \mathbb{R}^{C\times W\times H}$

First Pass: Cross-Shaped Attention
Compute query, key, and value using 1×1 convolutions:
    $Q \leftarrow \text{Conv}_{1\times1}(\overline{X})$
    $K \leftarrow \text{Conv}_{1\times1}(\overline{X})$
    $V \leftarrow \text{Conv}_{1\times1}(\overline{X})$
Construct affinity matrix D:
  for each spatial position u in Q **do**
    for each aligned position i **do**
      $d_{i,u} \leftarrow Q_u \cdot \Omega^{\top}_{i,u}$
    end for
  end for
Normalize using softmax to obtain attention map:
    $A \leftarrow \text{softmax}(D)$;
Aggregate contextual information:
  for each position u in $\overline{X}$ **do**
    $\overline{X}'_u \leftarrow \sum_{i\in|\Phi_u|} A_{i,u}\Phi_{i,u} + \overline{X}_u$
  end for
Second Pass: Contextual Propagation
Compute new attention map A' using $\overline{X}'$ following steps 5–16.
Update feature map $\overline{X}''$:
  for each position u in $\overline{X}'$ **do**
    if u and θ are in the same row or column **then**
      $\overline{X}''_u \leftarrow f(A', u, \theta)\cdot\overline{X}_\theta \cdot [f(A,u,\theta)+1]$
    else
      $\overline{X}''_u \leftarrow \sum f(A', A)\cdot\overline{X}_\theta$
  end for
Return $\overline{X}''$

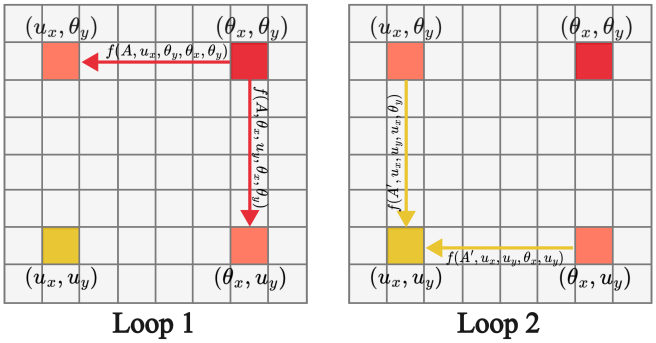


Fig. 4. After two iterations of the cross-shaped attention mechanism, each neuron is capable of fully acquiring dense contextual information from all other neurons.

branches, such as the two branches illustrated in Fig. 3, are merged using element-wise addition:

$$E = \tilde{E} + \overline{E} \tag{5}$$

Subsequently, we integrate global feature by using global average pooling to generate channel-specific statistics $s \in \mathbb{R}^C$. In particular, the c-th component of s is determined by summing E across the spatial dimensions $H \times W$:

$$s_c = Pooling(E_c) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} E_c(i,j). \tag{6}$$

Furthermore, we generate a streamlined feature vector $z \in \mathbb{R}^{d\times 1}$ to facilitate accurate and adaptive selection. This is accomplished through a straightforward fully connected (fc) layer, which improves efficiency and reduces the dimensionality:

$$z = F_{fc}(s) = \delta(B(Ws)), \tag{7}$$

Here, δ refers to the ReLU activation function, B indicates Batch Normalization, and $W \in \mathbb{R}^{d\times C}$. To evaluate how d affects model efficiency, we utilize a reduction ratio r to set its value:

$$d = \max(C/r, L), \tag{8}$$

Here, L represents the smallest value of d, with $L = 32$ commonly used in our experiments.

4) Selection: A soft attention mechanism is employed to dynamically select information across different spatial scales, guided by a compact feature descriptor z. This selection process computes the attention weights a_c and b_c for each channel c by applying the soft-max function to the values along the channel dimension. The attention weights are defined as follows:

$$a_c = \frac{e^{A_c z}}{e^{A_c z} + e^{B_c z}}, \quad b_c = \frac{e^{B_c z}}{e^{A_c z} + e^{B_c z}}, \tag{9}$$

where A and B are matrices with dimensions $\mathbb{R}^{C\times d}$, and a and b are the soft attention vectors corresponding to the feature maps $\overline{E}$ and $\tilde{E}$, respectively. Specifically, $A_c \in \mathbb{R}^{1\times d}$ and $B_c \in \mathbb{R}^{1\times d}$ represent the c-th row of matrices A and B, respectively. The attention weights a_c and b_c satisfy $a_c + b_c = 1$ in the two-branch case, eliminating the need for matrix B.

The final feature map V is generated by applying these attention weights to the corresponding kernels:

$$V_c = a_c \cdot \overline{E}_c + b_c \cdot \tilde{E}_c, \quad a_c + b_c = 1, \tag{10}$$

The aggregated feature map V is then formed as:

$$V = [V_1, V_2, ..., V_C], \quad V_c \in \mathbb{R}^{H\times W}. \tag{11}$$

This process supports the two-branch case, and can be easily extended to handle more branches by generalizing the attention weights and combining feature maps accordingly.

E. Primary-Secondary Modality Discrepancy Learning Module

PS-MDL Module evaluate the discrepancies among modalities, it is essential to first identify the primary modal and auxiliary modalities. In this work, we fuse the three modalities (t, a, v) and apply the CS-DSA module for global context extraction, resulting in a fourth modal—the fusion modal $\hat{X}_f \in \mathbb{R}^{c\times w\times H}$. By default, this fusion modal is treated as the primary modal, while the original three serve as auxiliary modalities (ablation studies demonstrate the performance

impact when other modalities are designated as the primary one):

$$\begin{aligned} X_f &= concat(X_t, X_a, X_v), \\ \hat{X}_f &= CS - DSA(X_f), \end{aligned} \tag{12}$$

To capture inconsistencies between the primary and auxiliary modalities, we employ a cross-attention (CA) mechanism that facilitates interaction among the auxiliary modalities:

$$\begin{aligned} \hat{S}_v &= CA(\hat{X}_a \to \hat{X}_v) \oplus CA(\hat{X}_t \to \hat{X}_v), \\ \hat{S}_a &= CA(\hat{X}_t \to \hat{X}_a) \oplus CA(\hat{X}_v \to \hat{X}_a), \\ \hat{S}_t &= CA(\hat{X}_a \to \hat{X}_t) \oplus CA(\hat{X}_v \to \hat{X}_t). \end{aligned} \tag{13}$$

Moreover, we also can calculate the fusion modal representations by cross-attention (CA) mechanism and self-attention (SA):

$$\begin{aligned} &\hat{S}_f = CA(\hat{X}_a \to \hat{X}_f) \oplus CA(\hat{X}_t \to \hat{X}_f) \oplus CA(\hat{X}_v \to \hat{X}_f), \\ &\hat{S}_f = SA(\hat{S}_f), \end{aligned} \tag{14}$$

Finally, the multimodal discrepancy representation is obtained by concatenating the auxiliary modalities $\hat{S}_v$, $\hat{S}_a$, and $\hat{S}_t$ with the primary modal $\hat{S}_f$:

$$O = W_1\hat{S}_v \oplus W_2\hat{S}_a \oplus W_3\hat{S}_t \oplus \hat{S}_f. \tag{15}$$

Here, W_1, W_2, and W_3 are trainable weight parameters that are learned by the modalities themselves to regulate the amount of auxiliary information to be extracted.

F. Output Layer

The task of multimodal sentiment analysis involves predicting the label y. Consequently, the final modal discrepancy representation is passed through a fully connected layer with a softmax activation function to produce a probability distribution y within the decision space of these tasks:

$$y = softmax(WO + b). \tag{16}$$

where Wo and bo are trainable parameters.

IV. Experiment

In this section, we will provide a detailed introduction to the datasets, baselines, Evaluation Metrics, and parameter settings used in our work.

A. Datasets

To evaluate the performance of our SSGDL model, we utilize three well-established benchmark datasets: CMU-MOSI [37], CMU-MOSEI [38], and CH-SIMS [39].. The purpose to choose these dataset because they are commonly used for multimodal sentiment analysis (MSA). Additionally, we incorporate an additional dataset, AVEC2019 [40], to verify the robustness of the proposed method. This, in turn, sheds light on the generalization capabilities and cross-domain performance of the proposed method. Below is a detailed summary of the datasets, including their training, validation, and testing set distributions, as outlined in Table I:

- CMU-MOSI: This dataset comprises text, visual, and acoustic data from 93 YouTube movie review videos, segmented into 2,199 parts. Each segment is annotated with a sentiment intensity score ranging from -3 to 3. The dataset is divided into 1,284 segments for training, 229 for validation, and 686 for testing.
- CMU-MOSEI: This larger-scale dataset includes 23,453 annotated video segments from various online platforms. It covers 250 topics and features 1,000 different speakers. Each segment is labelled with sentiment intensity scores from -3 to 3 and includes annotations for six basic emotions. This dual labelling supports both sentiment and emotion recognition tasks.
- CH-SIMS: Designed for sentiment analysis in a Chinese context, CH-SIMS features 2,281 utterance-level video segments from 60 diverse video sources, such as movies, TV dramas, and variety shows. Each segment is annotated with sentiment intensity scores from -1 (highly negative) to 1 (highly positive). The dataset is split into 1,368 segments for training, 456 for validation, and 457 for testing. Although it provides both multimodal and unimodal annotations, our study focuses on the multimodal annotations.
- AVEC2019: The AVEC2019 dataset is aimed at multimodal depression detection and includes audiovisual recordings from clinical interviews with a virtual agent and human interaction. Each sample is annotated with PHQ-8 scores ranging from 0 to 24, indicating the severity of depression. The dataset contains 163 training samples, 56 validation samples, and 56 test samples. For acoustic features, we use Mel-frequency cepstral coefficients (MFCC), and for visual features, we use facial action units.

TABLE I
SPLITTING RESULTS OF THE CMU-MOSI, CMU-MOSEI, CH-SIMS AND AVEC2019 DATASETS.

Datasets	Train	Valid	Test	All
CMU-MOSI [37]	1284	299	686	2199
CMU-MOSEI [38]	16326	1871	4659	22856
CH-SIMS [39]	1368	456	457	2281
AVEC2019 [40]	163	56	56	275

B. Baseline

To validate the effectiveness of our SSGDL model across both tasks, we compare our experimental results with the accuracy and performance of several state-of-the-art methods in sentiment analysis.

1) *Multimodal Sentiment Analysis*

TFN [24] TFN is a tensor-based approach that leverages the Cartesian product to derive a holistic representation of the involved modalities. It achieves this by employing a modal embedding subnetwork to learn intra-modality dynamics and a novel fusion method called tensor fusion to capture inter-modal interactions.

LMF [41] LMF discards traditional alignment methods for different modalities and instead employs stacked Transformers to expand the available temporal frames for alignment.

MulT [23] MulT employs directional pairwise cross-modal attentions to facilitate interactions between modalities. This is achieved by translating information from one modal to another and vice versa.

MFM [42] MFM is a multimodal factorization model that decomposes characterization factors using multimodal discriminant factors and modal-specific generation factors.

MFN [43] MFN utilizes LSTM-related structures to simultaneously process temporal information from three modalities.

Self-MM [22] Self-MM jointly trains multimodal and unimodal tasks using both multimodal labels and generated unimodal labels. This approach facilitates learning similarities and differences between modalities effectively.

MNT [27] MNT employs the self-attention mechanism of the Transformer to process cross-modal information, utilizing various normalization operations.

TETFN [44] The Text-Enhanced Transformer Fusion Network (TETFN) excels at creating cohesive multimodal representations by focusing on pairwise cross-modal interactions driven by text and highlighting the differences between modalities through the use of unimodal labels.

C-MIB [45] C-MIB leverages the information bottleneck concept, aiming to optimize the mutual information shared between unimodal and multimodal representations with their corresponding targets. It simultaneously restricts the mutual information between these representations and their inputs, guiding the model to learn an efficient and non-redundant multimodal representation.

MCL [46] The MCL approach leverages prior knowledge to uncover correlations among different modalities. It creates sets of positive and negative samples from the same instance and distinct instances, respectively. By utilizing a weak predictor to discern relationships within these sets, MCL helps the model to link unimodal features and identify commonalities across modalities.

TSST [47] TSST breaks down the fusion process into two distinct phases. Each phase is dedicated to capturing interactions between unimodal signals and the interaction information within fused representations, thereby enhancing the communication across different modalities.

TSCL-FHFN [48] TSCL-FHFN introduces an attention-driven directional cross-modal transformer that enables one modal to draw information from another, facilitating the acquisition of complementary data between them. FHFN then employs a low-rank tensor fusion strategy to reinforce the learning of interactions across multiple modalities.

MMML [58] MMML is a state-of-the-art model that combines audio and text signals using a transformer-based fusion network. It employs multi-loss training to optimize the performance of individual modalities and the overall fusion network.

2) *Multimodal Depression Detection*

Baseline [49] The baseline approach utilizes late fusion by averaging the final predictions from all involved modalities.

Adaptive Fusion Transformer [50] Sun et al. propose the Adaptive Fusion Transformer networks to dynamically combine the final predictions.

EF [51] EF incorporates straightforward linguistic and word-duration features to assess depression levels.

Bert-CNN & Gated-CNN [52] Bert-CNN & Gated-CNN are designed with gating mechanisms to integrate information from textual, acoustic, and visual modalities.

Multi-scale Temporal Dilated CNN [53] This method employs dilated CNNs to extract multimodal features, expanding the receptive field to handle longer sequences.

Hierarchical BiLSTM [54] Hierarchical BiLSTM utilizes a hierarchical BiLSTM structure to capture sequential data in a pyramid-like fashion.

TensorFormer [55] TensorFormer is a tensor-based Transformer framework for multimodal data, considering interactions among all relevant modalities.

C. Experimental Settings

Our model is optimized using the Adam optimizer, with an initial learning rate of 5×10^{-5} for RoBerta, and 1×10^{-3} for other parameters. We experiment with mini-batch sizes of 16, 32, and 64, adjust the LSTM hidden layer sizes to 32, 64, or 128, and set the kernel size for Conv1D to 3. The sequence length for visual features k is maintained at 50, and the number of attention heads for text-based multi-head attention is configured to 5.

D. Assessment Metrics

The experiment adopts four evaluation metrics to assess the performance of our model in multimodal sentiment analysis tasks. The specific metrics are as follows:

Mean Absolute Error (MAE) Measures the average absolute difference between the forecasted and actual values, disregarding the direction of the errors.

Binary Classification Accuracy (Acc-2) Evaluated in two scenarios: one comparing negative with non-negative (including zero) and another contrasting negative with positive (excluding zero).

F1 Score (F1) Assessed similarly in two contexts: negative versus non-negative (including zero) and negative versus positive (excluding zero).

Pearson Correlation Coefficient (Corr) Gauges the strength and direction of the linear relationship between the predicted values and the actual outcomes of the samples.

V. Results and Analysis

A. Quantitative Analysis

1) *Multimodal Sentiment Analysis*

Comparisons on CMU-MOSI and CMU-MOSEI Datasets. Table II showcases a performance comparison between our SSGDL model and several leading baseline models on the MOSI and MOSEI datasets. For each metric on these datasets, two sets of results are provided. The values to the left of the '/' represent the model's performance when zero-label samples are included in the dataset (has-0), while the

TABLE II
RESULTS ON THE CMU-MOSI AND CMU-MOSEI DATASETS. FOR EACH METRIC ON THESE DATASETS, TWO SETS OF RESULTS ARE REPORTED: VALUES ON THE LEFT OF THE '/' REPRESENT PERFORMANCE WHEN ZERO-LABEL SAMPLES ARE INCLUDED (HAS-0), WHILE VALUES ON THE RIGHT REFLECT PERFORMANCE WHEN THOSE SAMPLES ARE EXCLUDED (NON-0). IN THIS CONTEXT, THE SYMBOL "↑" INDICATES THAT HIGHER VALUES CORRESPOND TO BETTER PERFORMANCE, WHILE THE SYMBOL "↓ " DENOTES THAT LOWER VALUES INDICATE BETTER PERFORMANCE. THE SYMBOL "−" SIGNIFIES THAT THE RELEVANT VALUE IS NOT PROVIDED IN THE CORRESPONDING REFERENCE. THE BEST PERFORMANCE FOR EACH METRIC IS HIGHLIGHTED IN **BOLD**.

Models	CMU-MOSI				CMU-MOSEI			
	Acc_2 ↑	F1↑	MAE↓	Corr↑	Acc_2 ↑	F1↑	MAE↓	Corr↑
TFN [20]	79.15/80.95	79.03/80.9	0.933	0.672	79.96/81.38	80.2/81.45	0.913	0.686
LMF [13]	77.26/78.51	77.19/78.5	0.956	0.628	78.42/80.53	78.36/79.68	0.875	0.65
MulT [23]	78.28/79.73	78.3/79.81	0.908	0.696	79.5/80.64	80.3/81.54	0.865	0.715
MFM [24]	77.11/77.74	77.17/77.87	0.978	0.652	78.31/79.5	79.3/79.7	0.902	0.72
MFN [14]	77.2/78.81	76.82/78.6	0.902	0.681	78.92/79.31	78.8/79.54	0.843	0.726
Self-MM [17]	82.83/85.68	82.75/85.79	0.845	0.79	83.68/85.68	82.75/85.79	0.845	0.74
MNT [18]	-/84.5	-/85.74	0.857	0.782	-/84.72	-/85.6	0.77	0.728
TETFN [37]	85.36/85.58	85.16/85.43	0.612	0.834	85.68/85.9	85.78/86.23	0.623	0.765
C-MIB [61]	-/84.12	-/84.23	0.684	0.83	-/85.4	-/85.57	0.638	0.77
MCL [50]	-/83.4	-/83.7	0.796	0.788	-/84	-/86.3	0.73	0.79
TSST [23]	85.73/85.89	85.23/85.49	0.649	0.85	86.5/86.2	86.45/86.69	0.586	0.76
FHFN [27]	85.79/86.21	85.98/86.46	0.62	0.842	85.9/86.36	85.62/86.8	0.57	0.785
MMML [58]	85.91/88.16	85.85/88.15	0.643	0.838	86.32/86.73	86.23/86.49	0.517	0.791
Our Model	**89.17/89.34**	**90.32/89.84**	**0.58**	**0.877**	**88.86/89.04**	**89.54/89.35**	**0.587**	**0.872**

values to the right indicate performance when those samples are excluded (non-0). The best results for each metric are emphasized in bold.

From the experimental results, several key insights emerge: (1) Superior Performance on CMU-MOSI: Our proposed SSGDL method outperforms existing approaches on the CMU-MOSI dataset. Specifically, when compared to the recent state-of-the-art FHFN approach, our model exhibits significant improvements of 3.38%/3.13% in Acc2 and 4.34%/3.38% in F1 scores, respectively. performance gain can be attributed to the Cross-Shaped Dynamic Scale Attention (CS-DSA) module, which dynamically adjusts the receptive field scale based on the input stimulus. By capturing fine-grained intra-modal correlations and aggregating global contextual information at varying scales, the CS-DSA module enables the model to effectively handle the diverse and complex emotional expressions present in multimodal data.

(2) Strong Performance on CMU-MOSEI: On the larger MSA benchmark dataset, CMU-MOSEI, our method also demonstrates considerable gains across most evaluation metrics. Notably, it surpasses the next-best method TSST by 2.36%/2.84% in Acc2 and 3.09%/2.66% in F1, respectively. Additionally, our approach achieves the highest performance in regression accuracy metrics, including MAE and Corr. This success is largely due to the Primary-Secondary modal Discrepancy Learning (PS-MDL) module, which leverages cross-attention and self-attention mechanisms to capture discrepancies between the primary (fused) modal and auxiliary modalities (text, acoustic, and visual). By hierarchically integrating these modalities, the PS-MDL module ensures that the model effectively utilizes the complementarity and uniqueness of each modal, leading to a more nuanced understanding of emotional content.

(3) Combined Strengths of CS-DSA and PS-MDL: The superior performance of SSGDL across both datasets underscores the efficacy of combining global context extraction (via CS-DSA) and inter-modal discrepancy learning (via PS-MDL). The CS-DSA module's ability to adaptively capture global and local correlations, combined with the PS-MDL module's focus on inter-modal discrepancies, enables the model to better handle both objective and subjective ambiguities in sentiment analysis and depression detection. These results demonstrate that the SSGDL framework provides a more comprehensive and accurate representation of emotional content, thereby setting a new benchmark for multimodal affective computing.

However, it is also observed that the improvement of our SSGDL method on the CMU-MOSEI dataset is somewhat less pronounced compared to its performance on the CMU-MOSI dataset. We hypothesize that this may be attributed to the inherent diversity and complexity of the CMU-MOSEI dataset. As one of the largest benchmark datasets for MSA tasks, CMU-MOSEI encompasses a wider range of topics, speakers, and emotional expressions, making it significantly more intricate than CMU-MOSI. This added complexity introduces a greater imbalance between fine-grained and coarse-grained metrics, presenting more challenging conditions for sentiment analysis models to effectively capture and generalize emotional patterns.

Comparisons on CH-SIMS Dataset. On the CH-SIMS dataset, we evaluate binary accuracy (Acc2), F1 score, Pearson correlation coefficient, and MAE. For all metrics except MAE, higher values indicate better performance. The CH-SIMS dataset stands out due to its use of Chinese text, distinguishing it from the other two datasets that focus on a limited number of English-language works. The performance results are presented in Table III.

From the comparisons in Table III, we can derive several key insights: (1) Our method shows notable improvements on this Chinese-language dataset. Specifically, the proposed SSGDL model surpasses the second-best approach, TSST, with gains of 1.5% in multi-class classification and 1.02% in binary

classification metrics. This underscores the effectiveness of our approach across different languages, even with varying feature extraction techniques. (2) It is also important to highlight the relative lack of sentiment analysis research specifically tailored to the Chinese language in multimodal datasets like CH-SIMS. Compared to existing state-of-the-art methods, our SSGDL model achieves significant performance enhancements on this dataset, demonstrating its ability to address sentiment analysis challenges by effectively managing both objective and subjective ambiguities in human emotion analysis.

TABLE III
RESULTS ON THE CH-SIMS DATASET.

Model	Acc_2 ↑	F1↑	MAE↓	Corr↑
TFN [24]	74.48	74.37	0.79	0.485
MulT [23]	75.29	75.58	0.785	0.502
MFN [43]	76.2	76.35	0.782	0.52
Self-MM [22]	80.7	80.65	0.754	0.613
MNT [27]	77.94	78.15	0.774	0.7542
TETFN [44]	79.17	79.28	0.75	0.76
MCL [46]	81.88	81.9	0.764	0.78
TSST [47]	82.3	82.71	0.76	0.746
MMML [58]	82.93	82.9	0.332	0.733
Our Model	**83.8**	**83.73**	**0.759**	**0.803**

2) *Comparisons on AVEC2019 Dataset*

Table IV showcases the performance of our proposed SSGDL model on the publicly available AVEC2019 multimodal dataset for depression detection. Upon analysing the results, it is evident that our method demonstrates robust performance in detecting depression. Notably, the SSGDL model surpasses the second-best approach, TensorFormer, achieving a 0.041% increase in the CCC metric and a 0.52% decrease in RMSE. These findings highlight the sustained efficacy of our approach in the field of depression detection. Thus, the results proof the generalization capabilities and cross-domain performance of our proposed method.

TABLE IV
RESULTS ON THE AVEC2019 DATASET.

Model	CCC↑	RMSE↓
Baseline [49]	0.111	6.37
Adaptive Fusion Transformer [50]	0.443	5.61
EF [51]	0.344	-
Bert-and-Gated-CNN [52]	0.403	6.11
Multi-scale Temporal Dilated CNN [53]	0.430	4.39
Hierarchical BiLSTM [54]	0.442	5.50
TensorFormer [55]	0.493	4.31
Our Model	**0.534**	**3.79**

B. Case Study

Furthermore, to evaluate our model's predictive capabilities, we performed case studies using three selected video clips from the CMU-MOSI dataset. Fig. 5 displays the sentiment prediction outcomes for each clip, including corresponding text, acoustic, and visual data, alongside the actual label values and the model's predictions. Here, negative values correspond to negative sentiments, while positive values signify positive sentiments. The figure demonstrates a strong alignment between the predicted values and the true labels, providing compelling evidence of the model's accuracy. These results, combined with factors such as the incorporation of global and spatial data, the deployment of a sophisticated feature fusion strategy, and the success observed in the case studies, significantly bolster the model's overall performance and confirm its effectiveness.

In Clip A, the model correctly predicts a positive sentiment despite the presence of conflicting cues in the acoustic modal. This accurate prediction can be attributed to the Cross-Shaped Dynamic Scale Attention (CS-DSA) module's ability to dynamically adjust the receptive field scale based on the input stimulus. By focusing on the most relevant contextual information in the visual and textual modalities, the CS-DSA module effectively captures both global and local correlations. The cross-shaped attention mechanism ensures that the model aggregates comprehensive contextual information, leading to a more accurate sentiment prediction even in the presence of conflicting acoustic cues. In Clip B, the model successfully identifies a negative sentiment by leveraging the Primary-Secondary Modality Discrepancy Learning (PS-MDL) module's ability to capture discrepancies between the fused modal and the acoustic modal. The cross-attention mechanism highlights subtle differences in tone and pitch, which are critical for detecting negative emotions. This demonstrates how the PS-MDL module's hierarchical integration of modalities enables the model to effectively utilize inter-modal discrepancies for sentiment analysis. By emphasizing the unique contributions of each modal, the PS-MDL module ensures a more nuanced understanding of emotional content. In Clip C, the model accurately predicts a negative sentiment by balancing the contributions of all modalities. The CS-DSA module's dynamic scale selection allows the model to adaptively capture relevant contextual information, while the PS-MDL module's cross-attention mechanism ensures that no single modal dominates the prediction. This balanced approach highlights the model's ability to handle complex and ambiguous scenarios, where multiple modalities may provide conflicting or complementary information.

These case studies demonstrate that the SSGDL model's ability to adaptively capture global context (via CS-DSA) and inter-modal discrepancies (via PS-MDL) enables it to accurately predict sentiment even in complex and ambiguous scenarios. The integration of these mechanisms ensures that the model can effectively leverage the strengths of each modal while mitigating the impact of conflicting cues, leading to robust and accurate sentiment analysis.

C. Ablation Study

The proposed model comprises two key components: CS-DSA and PS-MDL. To gain a deeper understanding of their respective contributions, we conducted a comprehensive ablation study on the CMU-MOSI dataset, involving four key analyses. First, we systematically removed each internal component while preserving the overall model structure to assess its individual impact on performance. Second, we evaluated the influence of each modality (text, visual, and acoustic) as the

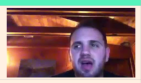
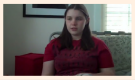
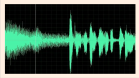
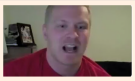
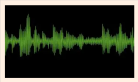
Text	Video	Audio	Truth	Prediction
I think is really nice			2.2	2.22
I was definitely expecting this to be a good movie			1.5	1.49
I couldnt stand UP And I really hated Cars			-2.6	-2.61

Fig. 5. Case Study Result. The samples in our dataset were labelled on a scale ranging [-3, 3]. Values greater than 0 indicate positive sentiment, with -3 representing the most positive sentiment. Conversely, values less than 0 indicate negative sentiment, with 3 representing the most negative sentiment.

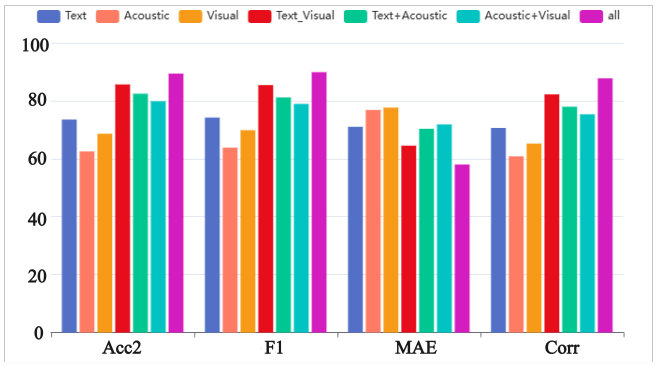


Fig. 6. The ablation study results for utilizing individual modalities or combining multiple modalities.

primary modality to analyse modality-specific discrepancies and assess the effectiveness of using the fusion modality as the primary one. These experiments provide critical insights into the role of each module, their relative importance, and how they contribute to the overall architecture. Additionally, we examined unimodal and multimodal configurations to demonstrate the advantages of multimodal fusion for sentiment analysis and depression detection. Finally, we further validated the effectiveness of CS-DSA by replacing it with multiple benchmark fusion techniques. In these experiments, CS-DSA refers to the Cross-Shaped Dynamic Scale Attention module, PS-MDL denotes the Primary-Secondary Modality Discrepancy Learning module, where module removals are indicated by "−" and combined configurations are represented by "+".

TABLE V
ABLATION EXPERIMENT RESULTS FOR EACH COMPONENT.

Model	Acc_2 ↑	F1↑	MAE↓	Corr↑
non	81.52/82.4	81.6/82.75	0.842	0.601
- CS-DSA	86.38/86.84	87.02/87.75	0.741	0.748
- PS-MDL	87.24/87.55	86.9/87	0.81	0.72
+ All Module	**89.17/89.34**	**90.32/89.84**	**0.58**	**0.877**

1) The experimental results in Table V demonstrate that removing any internal component leads to varying degrees of performance degradation, highlighting the critical role each component plays in enhancing the overall effectiveness of the model. When all major modules are removed, the performance drops sharply—up to 7.65% in Acc2 and 8.74% in F1—confirming that these components are not merely additive enhancements, but essential to the model's representational capacity. Specifically, when the CS-DSA module is removed, the Acc2 and F1 scores decrease by 2.79%/2.5%. This reflects the importance of scale-adaptive attention, which enables the model to flexibly integrate both local and global contextual cues. Without this mechanism, the model is forced to attend to fixed receptive fields, which may either miss long-range sentiment evolution or dilute sharp local emotional triggers—especially in temporally misaligned modalities like speech and vision. Similarly, when the PS-MDL module is excluded, Acc2 and F1 drop by 1.93%/1.79% and 3.42%/2.84%, respectively. This demonstrates the importance of capturing inter-modal discrepancies through cross-attention and self-attention mechanisms. The PS-MDL module's hierarchical integration of auxiliary modalities ensures that the model can leverage the unique contributions of each modal, leading to a more robust fused representation.

Together, these findings validate our hypothesis that accurate affective understanding requires both context-sensitive aggregation (CS-DSA) and discrepancy-aware alignment (PS-MDL). Their combined contribution forms a dual mechanism: one that ensures comprehensive context modelling, and another that guards against over-reliance on potentially misleading modalities. These insights go beyond numerical gains—they reflect a cognitive-inspired architecture design that is both interpretable and generalizable.

TABLE VI
PERFORMANCE COMPARISON WITH LARGE VISION MODELS (IMAGE AND TEXT ONLY).

Model	Acc_2 ↑	F1↑	MAE↓	Corr↑
LLava [59]	85.31	86.12	0.71	0.82
LLava Video [60]	86.03	86.78	0.69	0.83
LLama [61]	84.72	85.51	0.73	0.81
Our Model	**87.14**	**87.7**	**0.645**	**0.842**

TABLE VII
RESULTS OF ABLATION EXPERIMENTS WHEN EACH SINGLE MODE IS USED AS THE PRIMARY MODAL.

Model	Acc_2 ↑	F1↑	MAE↓	Corr↑
Text	87.59/87.58	87.8/87.77	0.671	0.862
Audio	86.63/86.8	87.32/87.42	0.69	0.841
Visual	87.24/87.16	87.51/87.33	0.679	0.854
Fusion	**89.17/89.34**	**90.32/89.84**	**0.58**	**0.877**

2) To further demonstrate the importance of multimodal data in sentiment analysis tasks, we conducted experiments with various data configurations: text-only (T), audio (A), video (V), video + audio (V + A), video + text (V + T), and audio + text (A + T). These experiments were designed to evaluate the complementary contributions of visual and acoustic modalities to textual information. The results, presented in Fig. 6, indicate that performance degrades across

all partial-modality configurations, reaffirming that sentiment expression is inherently multimodal. Notably, the audio-only setup yields the lowest performance, which can be attributed to the inherent variability and noise in acoustic signals, making them less reliable for consistent emotional inference. In contrast, configurations including text consistently perform better—particularly text + video (T + V)—as text typically carries explicit semantic sentiment, while facial expressions serve as rich affective complements. This pattern confirms that different modalities offer non-redundant emotional cues, and cross-modal alignment enhances robustness in interpretation, especially under ambiguous or implicit emotional scenarios.

To comprehensively evaluate the robustness of our model on datasets with only image and text modalities, we introduced LLava [59], LLava Video [60], and LLama [61] models for comparison. These models are state-of-the-art in handling visual and textual modalities, providing a strong benchmark for our model.

The results of the ablation experiments without audio inputs are presented in Table.VI. Our model demonstrates superior performance compared to LLava, LLava Video, and LLama in terms of accuracy (Acc2) and F1 score. Specifically, our model outperforming LLava (Acc2: 85.31%, F1: 86.12%), LLava Video (Acc2: 86.03%, F1: 86.78%), and LLama (Acc2: 84.72%, F1: 85.51%). This performance advantage suggests two key insights. One one hand, our architecture not only fuses modalities but models their relative reliability through the PS-MDL module, allowing the system to focus on modality-consistent features and ignore irrelevant or misleading signals. On the other hand, the CS-DSA module adaptively adjusts its attention scale, enabling fine-grained alignment between visual and textual information over varying spatial and temporal scopes, which standard vision-language models (trained primarily for static grounding or generation tasks) often lack.

These findings reinforce the importance of task-specific multimodal alignment mechanisms over general-purpose fusion, and demonstrate our model's ability to outperform larger pre-trained models in affective reasoning through more targeted and cognitively aligned designs.

3) To investigate the effect of primary modality selection within the PS-MDL framework, we conducted a series of ablation experiments where each individual modality—text, audio, or video—was treated as the primary input, instead of the fused multimodal representation. The results (Table.VII) show that this substitution consistently leads to performance degradation, with accuracy and F1 drops of up to 1.5–2.5% depending on the dataset and modality.

This decline highlights a key insight: while individual modalities contain rich information, they are also prone to modality-specific bias or noise. For example, textual inputs may dominate in explicit sentiment cues but fail to reflect tonal sarcasm or facial dissonance. Visual cues may be expressive yet ambiguous without linguistic context. Using a single modality as the primary reference forces the model to interpret discrepancies relative to a potentially unstable or biased baseline. In contrast, the fused modality representation serves as a more reliable semantic anchor. It integrates shared affective signals across modalities, suppresses modality-specific noise, and offers a consensus-level baseline for discrepancy modelling. This design mirrors how humans interpret emotion: we often form a general impression based on combined cues before noticing incongruence.

These results confirm that treating the fused modality as primary enhances semantic stability, strengthens cross-modal calibration, and enables more robust handling of emotionally ambiguous or contradictory inputs. It also reinforces the design philosophy behind PS-MDL: that affective reasoning benefits from discrepancy detection around a central, semantically averaged core.

4) To validate the effectiveness of the CS-DSA module, we replaced it with several state-of-the-art fusion methods, including Multi-head Cross Attention, Low-Rank Fusion, and FHFN, keeping the rest of the model unchanged. As shown in Table.VIII, CS-DSA achieves the best results across all metrics on the MOSI dataset, outperforming the best baseline (FHFN) by over 2.3% in accuracy.

These improvements reflect more than just numerical gains. Traditional fusion methods often rely on static attention patterns, which struggle with temporal misalignment and diverse emotional cues. In contrast, CS-DSA introduces adaptive attention scaling and a cross-shaped structure, enabling the model to focus flexibly on both long-range dependencies (e.g., narrative sentiment flow) and localized cues (e.g., micro-expressions or tonal shifts).

This design allows more precise and context-sensitive multimodal integration, explaining its consistent advantage over fixed-fusion methods. The results highlight CS-DSA's suitability for affective reasoning tasks, where sentiment is both multi-faceted and temporally dynamic.

TABLE VIII
PERFORMANCE COMPARISON WITH DIFFERENT FUSION METHODS

Model	Acc_2 ↑	F1↑	MAE↓	Corr↑
MulT [23]	83.2/83.73	83.1/83.95	0.8	0.72
MFN [43]	81.8/82.45	81.94/82.73	0.853	0.694
Self-MM [22]	83.52/85.62	83.45/85.86	0.813	0.782
TETFN [44]	86.18/86.7	85.96/86.6	0.605	0.849
TSST [47]	86.5/86.16	86.18/86.84	0.622	0.86
FHFN [48]	86.8/87.36	86.5/87.42	0.61	0.85
Our Model	**89.17/89.34**	**90.32/89.84**	**0.58**	**0.877**

VI. LIMITATION

Although the proposed model effectively captures both intra-modal and cross-modal differences through the CS-DSA and PS-MDL modules, it does not specifically address the potential inconsistencies that may arise between bimodal data and the fusion of unimodal and bimodal inputs (e.g., text and audio). In practical sentiment analysis tasks, the integration of multiple modalities presents inherent challenges, particularly when the emotional expressions conveyed by different modalities conflict or diverge. This is especially true in complex and diverse scenarios, where inconsistencies in affective information across modalities may emerge.

The primary focus of this work, however, is on extracting global contextual information and learning the relationships

between modalities in multimodal sentiment analysis. Consequently, our efforts are concentrated on addressing the fusion challenges between unimodal and multimodal data. While the issue of cross-modal inconsistencies is certainly important, it is not the central concern of this study.

We assert that ensuring the effective fusion of unimodal and multimodal data to achieve accurate global context extraction should be prioritized in the initial phase of development. Once these foundational challenges are addressed, exploring the inconsistencies between bimodal data and between unimodal and bimodal inputs can become a valuable direction for future research. Although these inconsistencies may impact the model's performance in certain contexts, they do not fundamentally alter the core contributions of this paper. Therefore, while this limitation exists, it is considered a localized constraint that does not diminish the overall significance of the study.

VII. Conclusion and Future Work

This paper presents the Scale-Selectable Global Information and Discrepancy Learning Network (SSGDL), an advanced model designed for multimodal sentiment analysis. Drawing inspiration from neuroscience, where receptive field sizes in the visual cortex adapt dynamically to stimuli, our model utilizes a cross-shaped attention mechanism (CS-DSA) to autonomously determine scale sizes. This mechanism effectively captures both global and contextual information by incorporating processes such as division, cross-interaction, fusion, and selection, which enhances the model's ability to capture intra-modal details. To address the cross-modal discrepancies often neglected in prior research, we introduce the Primary-Secondary modal Discrepancy Learning (PS-MDL) module. In this setup, the fusion modal produced by CS-DSA serves as the primary modal, with other modalities acting as secondary. By leveraging cross-attention, the model learns and addresses discrepancies between these modalities, thereby improving its interpretation of emotional content and enriching the representation of emotions. Experimental evaluations reveal that our model performs competitively on benchmark datasets. Looking ahead, future work will focus on enhancing accuracy through multi-task learning and refining the architecture to achieve a more streamlined design.

References

[1] J. Li, B. Chiu, S. Shang, and L. Shao, *Neural Text Segmentation and Its Application to Sentiment Analysis*, *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 2, pp. 828–842, 2020.

[2] R. Ren and D. Wu, "An Innovative Sentiment Analysis to Measure Herd Behavior," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 10, pp. 3841-3851, Oct. 2020, doi: 10.1109/TSMC.2018.2864942.

[3] J. Yang, D. She, Y.-K. Lai, and M.-H. Yang, *Retrieving and Classifying Affective Images via Deep Metric Learning*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, pp. 491–498, 2018.

[4] S. Yang, L. Xing, Z. Chang, Y. Li, *et al.*, *Attention-Based Sentiment Region Importance and Relationship Analysis for Image Sentiment Recognition*, *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–14, 2022.

[5] E. Lieskovska, M. Jakubec, R. Jarina, M. Chmulič, *A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism*, *Electronics*, vol. 10, no. 10, pp. 1163, 2021.

[6] H. Zou, Y. Si, C. Chen, D. Rajan, E. S. Chng, *Speech Emotion Recognition with Co-Attention Based Multi-Level Acoustic Information*, in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 7367–7371, 2022.

[7] N. Xu, W. Mao, and G. Chen, *A co-memory network for multimodal sentiment analysis*, in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 929–932, 2018.

[8] S. Nemati, R. Rohani, and M. E. Basiri, *A hybrid latent space data fusion method for multimodal emotion recognition*, *IEEE Access*, vol. 7, pp. 172948–172964, 2019.

[9] Y. Yu, H. Lin, and J. Meng, *Visual and textual sentiment analysis of a microblog using deep convolutional neural networks*, *Algorithms*, vol. 9, no. 2, pp. 41, 2016.

[10] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, *Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data*, *Algorithms*, vol. 33, no. 1, pp. 7216–7223, 2019.

[11] A. Kumar, K. Srinivasan, and W. H. Cheng, *Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data*, *Journal of Computer Science and Technology*, vol. 57, no. 1, pp. 102141, 2020.

[12] A. Kumar, K. Srinivasan, and W. H. Cheng, *Multimodal hypergraph learning for microblog sentiment prediction*, in *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2015.

[13] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, *Multi-attention recurrent network for human communication comprehension*, in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[14] Y. Wu, Y. Zhao, X. Lu, B. Qin, Y. Wu, J. Sheng, and J. Li, *Modeling Incongruity Between Modalities for Multimodal Sarcasm Detection*, *IEEE MultiMedia*, vol. 28, no. 2, pp. 86–95, 2021.

[15] L. Sun, Z. Lian, B. Liu, and J. Tao, "Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, vol. 15, no. 1, pp. 309–325, 2024, doi: 10.1109/TAFFC.2023.3274829.

[16] H. Sun, Y.-W. Chen, and L. Lin, "A tensor-based multimodal transformer for multimodal sentiment analysis and depression detection," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2776–2786, 2023, doi: 10.1109/TAFFC.2022.3233070.

[17] A. Kumar, K. Srinivasan, and W. H. Cheng, *Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data*, *Journal of Computer Science and Technology*, vol. 57, no. 1, pp. 102141, 2020.

[18] S. A. Qureshi, G. Dias, M. Hasanuzzaman, and S. Saha, "Improving depression level estimation by concurrently learning emotion intensity," *IEEE Computational Intelligence Magazine*, vol. 15, no. 3, pp. 47–59, Aug. 2020.

[19] L. Zhu, M. Xu, Y. Bao, Y. Xu, and X. Kong, *Deep learning for aspect-based sentiment analysis: a review*, *PeerJ Comput. Sci*, vol. 8, p. 1044, 2022.

[20] R. Kapoor, M. Bhat, and N. Singh, *Recent advances in the discipline of text based affect recognition*, *Multimed Tools Appl*, pp. 48859–48893, 2024.

[21] H. Sun, Y. W. Chen, and L. Lin, "Tensorformer: A tensor-based multimodal transformer for multimodal sentiment analysis and depression detection," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2776–2786, 2022.

[22] Y. Wu, H. Xu, Z. Yuan, and J. Wu, *Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis*, *The Thirty-Fifth AAAI Conference on Artificial Intelligence(AAAI-21)*, 2021, pp. 10790-10797.

[23] 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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 6558–6569.

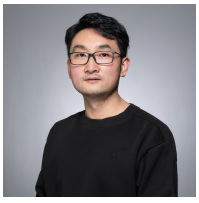
[24] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor Fusion Network for Multimodal Sentiment Analysis," 2017, arXiv:1707.07250.

[25] Zuhe Li, Zhenwei Huang, Yushan Pan, Jun Yu, Weihua Liu, Haoran Chen, Yiming Luo, Di Wu, Hao Wang, *Hierarchical denoising representation disentanglement and dual-channel cross-modal-context interaction for multimodal sentiment analysis*, *Expert Systems with Applications*, vol. 252, pp. 124236, 2024, Elsevier.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 1–11.

[27] J.-B. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, "A Transformer-Based Joint-Encoding for Emotion Recognition and Sentiment Analysis," in *Proceedings of the Association for Computational Linguistics (ACL)*, 2020, pp. 1–7.
[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, *An image is worth 16x16 words: Transformers for image recognition at scale*, *ICLR*, 2021.
[29] J. Yang, X. Dong, L. Liu, C. Zhang, J. Shen, and D. Yu, *Recurring the transformer for video action recognition*, in *the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14063–14073, 2022.
[30] H. Zhu, C. Cui, L. Deng, R. C. C. Cheung, and H. Yan, "Elastic Net Constraint-Based Tensor Model for High-Order Graph Matching," *IEEE Transactions on Cybernetics*, vol. 51, no. 8, pp. 4062–4074, Aug. 2021.
[31] C. Fan, K. Zhu, J. Tao, G. Yi, J. Xue, and Z. Lv, "Multi-level Contrastive Learning: Hierarchical Alleviation of Heterogeneity in Multimodal Sentiment Analysis," *IEEE Transactions on Affective Computing*, doi: 10.1109/TAFFC.2024.3423671.
[32] L. He, Z. Wang, L. Wang, and F. Li, *Multimodal Mutual Attention-Based Sentiment Analysis Framework Adapted to Complicated Contexts*, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7131–7143, Dec. 2023, doi: 10.1109/TCSVT.2023.3276075.
[33] W. Guo, J. Wang, and S. Wang, *Deep Multimodal Representation Learning: A Survey*, *IEEE Access*, vol. 7, pp. 63,373–63,394, 2019.
[34] H. Cheng, Z. Yang, X. Zhang, and Y. Yang, *Multimodal Sentiment Analysis Based on Attentional Temporal Convolutional Network and Multi-Layer Feature Fusion*, *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 3149–3163, Oct.–Dec. 2023, doi: 10.1109/TAFFC.2023.3265653.
[35] K. Kroenke and R. L. Spitzer, "The PHQ-9: A New Depression Diagnostic and Severity Measure," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 509–515, 2002.
[36] Z. Zhao, Q. Li, N. Cummins, B. Liu, H. Wang, J. Tao, and B. W. Schuller, "Hybrid Network Feature Extraction for Depression Assessment from Speech," in *Proceedings of INTERSPEECH*, 2020, pp. 4956–4960.
[37] E. Pincus, A. Zadeh, R. Zellers, and L.-P. Morency, *MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos*, arXiv:1606.06259, 2016.
[38] S. Poria, E. Cambria, A. Zadeh, P. Liang, and L.-P. Morency, *Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph*, in *56th Annual Meeting of the Association for Computational Linguistics*, pp. 2236–2246, 2018.
[39] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, and K. Yang, "CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Online, 2020, pp. 3718–3727.
[40] F. Ringeval *et al.*, "AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition," in *Proceedings of the 9th International Audio/Visual Emotion Challenge Workshop*, 2019, pp. 3–12.
[41] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, *Efficient low-rank multimodal fusion with modality-specific factors*, arXiv preprint arXiv:1806.00064, 2018.
[42] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, *Learning factorized multimodal representations*, arXiv preprint arXiv:1806.06176, 2018.
[43] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, *Memory fusion network for multi-view sequential learning*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.
[44] D. Wang, X. Guo, Y. Tian, J. Liu, L. Huo, and X. Luo, *TETFN: A Text Enhanced Transformer Fusion Network for Multimodal Sentiment Analysis*, *Pattern Recognition*, vol. 136, p. 109259, 2023.
[45] S. Mai, Y. Zeng, and H. Hu, *Multimodal Information Bottleneck: Learning Minimal Sufficient Unimodal and Multimodal Representations*, *IEEE Transactions on Multimedia*, vol. 25, pp. 4121–4134, 2022.
[46] S. Mai, Y. Sun, Y. Zeng, and H. Hu, *Excavating Multimodal Correlation for Representation Learning*, *Information Fusion*, vol. 91, pp. 542–555, 2023.
[47] G. Yi, C. Fan, J. Tao, Z. Lv, Z. Wen, G. Pei, and T. Li, *A Two-Stage Stacked Transformer Framework for Multimodal Sentiment Analysis*, *Intelligent Computing*, vol. 3, p. 0081, 2024.
[48] Y. Li, W. Weng, and C. Liu, *TSCL-FHFN: Two-Stage Contrastive Learning and Feature Hierarchical Fusion Network for Multimodal Sentiment Analysis*, *Neural Computing and Applications*, pp. 1–15, 2024.
[49] F. Ringeval *et al.*, "AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition," in *Proceedings of the 9th International Audio/Visual Emotion Challenge Workshop*, 2019, pp. 3–12.
[50] H. Sun *et al.*, "Multi-modal adaptive fusion transformer network for the estimation of depression level," *Sensors*, vol. 21, no. 14, 2021, Art. no. 4764.
[51] H. Kaya *et al.*, "Predicting depression and emotions in the crossroads of cultures, para-linguistics, and non-linguistics," in *Proceedings of the 9th International Audio/Visual Emotion Challenge Workshop*, 2019, pp. 27–35.
[52] M. Rodrigues Makiuchi, T. Warnita, K. Uto, and K. Shinoda, "Multimodal fusion of BERT-CNN and gated CNN representations for depression detection," in *Proceedings of the 9th International Audio/Visual Emotion Challenge Workshop*, 2019, pp. 55–63.
[53] W. Fan, Z. He, X. Xing, B. Cai, and W. Lu, "Multi-modality depression detection via multi-scale temporal dilated CNNs," in *Proceedings of the 9th International Audio/Visual Emotion Challenge Workshop*, 2019, pp. 73–80.
[54] S. Yin, C. Liang, H. Ding, and S. Wang, "A multi-modal hierarchical recurrent neural network for depression detection," in *Proceedings of the 9th International Audio/Visual Emotion Challenge Workshop*, 2019, pp. 65–71.
[55] H. Sun, Y. W. Chen, and L. Lin, "Tensorformer: A tensor-based multimodal transformer for multimodal sentiment analysis and depression detection," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2776–2786, 2022.
[56] M. Li, Z. Zhu, K. Li, and H. Pei, "Diversity and Balance: Multimodal Sentiment Analysis using Multimodal-Prefixed and Cross-Modal Attention," *IEEE Transactions on Affective Computing*, doi: 10.1109/TAFFC.2024.3430045.
[57] S. Wen *et al.*, "Memristive LSTM Network for Sentiment Analysis," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 3, pp. 1794-1804, Mar. 2021, doi: 10.1109/TSMC.2019.2906098.
[58] Z. Wu, Z. Gong, J. Koo, et al., "Multimodal multi-loss fusion network for sentiment analysis," in *Proc. 2024 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (Vol. 1: Long Papers)*, 2024, pp. 3588–3602.
[59] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Adv. Neural Inf. Process. Syst.*, vol. 36, 2024.
[60] B. Lin, Y. Ye, B. Zhu, et al., "Video-LLaVA: Learning united visual representation by alignment before projection," arXiv preprint arXiv:2311.10122, 2023.
[61] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., "LLaMA: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023.

Xiaojiang He is a PhD candidate at the University of Liverpool and is affiliated with Xi'an Jiaotong-Liverpool University, Suzhou, China. His research interests include Computer Vision, Sentiment Analysis, Human-computer Interaction, and Machine Learning.



Yushan Pan is an Assistant Professor at Xi'an Jiaotong-Liverpool University, Suzhou, China. He received his Ph.D. in Informatics from the University of Oslo in 2018. He was a Postdoctoral Fellow of Computer Engineering with the Faculty of Engineering, Norwegian University of Science and Technology, from 2018-2020. Previously, he was a tenured senior researcher at Norwegian Maritime Competence Center, affiliate with Norwegian University of Science and Technology, from 2020 to 2022. His research interests include Multimodal Sentiment Analysis, Affective computing detection, Human-computer Interaction.

Xinfei Guo received his Ph.D. in Computer Engineering from the University of Virginia, where he worked on cross-layer design approaches for reliability and low power. His research interests include hardware/software co-design for edge AI, machine learning-assisted chip design methodologies, and AI acceleration. Xinfei now is the Chief Associate Editor for IEEE Transactions on Vary Large Scale Integration Systems.

Zhijie Xu has served as the Head of the Department of Computer Science and Professor at Xi'an Jiaotong-Liverpool University since September 2024. Previously, he was a tenured professor and Dean at the University of Huddersfield, UK. With over 20 years of experience, Prof. Xu has supervised 18 PhD students and led numerous projects in fields such as machine vision, visual systems, digital twins, and edge computing. His research has received funding from the UK government, EU, and various companies, totaling over £2 million. Prof. Xu is a senior member of IEEE, IET, BCS, and HEA, and has made significant contributions to virtual reality and automation. He has served as a chair at multiple international conferences and as an editorial board member for renowned journals.

Chenguang Yang (Fellow, IEEE) received the B.Eng. degree in measurement and control from Northwestern Polytechnical University, Xi'an, China, in 2005, and the Ph.D. degree in control engineering from the National University of Singapore, Singapore, in 2010. He was a Postdoctoral Fellow of Human Robotics with Imperial College London, London, U.K., from 2009 to 2010. His research interest lie in human–robot interaction and intelligent system design. He is the Corresponding Co-Chair of IEEE Technical Committee on Collaborative Automation for Flexible Manufacturing.