Towards Multimodal Sentiment Analysis via Hierarchical Correlation Modeling with Semantic Distribution Constraints

Qinfu Xu[1], Yiwei Wei[2*], Chunlei Wu[1], Leiquan Wang[1], Shaozu Yuan[3], Jie Wu[1], Jing Lu[1], Hengyang Zhou[2]

[1]Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China)
[2]China University of Petroleum (Beijing) at Karamay, [3] JD AI Research
xqfupc@163.com, weiyiwei@cupk.edu.cn

Abstract

Sentiment analysis is rapidly advancing by utilizing various data modalities (e.g., text, video, and audio). However, most existing techniques only learn the atomic-level features that reflect strong correlations, while ignoring more complex compositions in multimodal data. Moreover, they also neglected the incongruity in semantic distribution among modalities. In light of this, we introduce a novel Hierarchical Correlation Modeling Network (HCMNet), which enhances the multimodal sentiment analysis by exploring both the atomic-level correlations based on dynamic attention reasoning and the composition-level correlations through topological graph reasoning. In addition, we also alleviate the impact of distributional inconsistencies between modalities from both atomic-level and composition-level perspectives. Specifically, we first design an atomic-level contrastive loss that constrains the semantic distribution across modalities to mitigate the atomic-level inconsistency. Then, we design a graph optimal transport module that integrates transport flows with different graphs to constrain the composition-level semantic distribution, thus reducing the inconsistency of compositional nodes. Experiments on three public benchmark datasets have demonstrated the superiority of the proposed model over the state-of-the-art methods.

Introduction

Multimodal Sentiment Analysis (MSA), a challenging but significant research topic, has gained increasing attention and more scientific efforts owing to its facility to convey emotions and views of individuals (Veltmeijer, Gerritsen, and Hindriks 2021; Zhang, He, and Lu 2019; Song et al. 2021). It aims to learn emotional information from mixed data containing multiple modalities (e.g., text, video, and audio) and make judgments based on psychological categorization. Current methods (Hu et al. 2021; Ma et al. 2022; Hu et al. 2022; Shi and Huang 2023; Tu et al. 2024) that aim to learn efficient emotional representations for MSA heavily rely on the hypothesis that different modalities exhibit evident correlations and design diversified fusion methods, which facilitates the cross-modal representation for final prediction. Some of the previous approaches (Hu et al. 2021; Chudasama et al. 2022; Shi and Huang 2023) employ advanced attention mechanisms to integrate strong correlations for emotion modeling. Other studies (Hu et al. 2022; Yang et al. 2023) focus on designing unified solutions for dual granularity emotion recognition. With the equipment of fine-grained multimodal content modeling, they have constantly promoted correlation mining for MSA task.

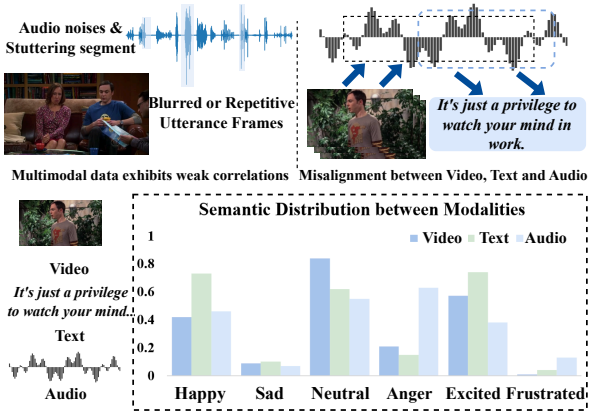

Figure 1: Illustration of hierarchical correlation modeling and the semantic distributions. **Up:** shows the complexity of multimodal correlations. Previous MSA methods tend to learn atomic-level correlations between modalities, ignoring the composition-level modeling that aims to capture weak but vital correlations. **Down:** shows the semantic distribution of a case obtained by unimodal classification, which indicates the incongruity between unimodal semantics.

Although promising, they still suffer from two limitations. On the one hand, most of the existing methods only consider the atomic-level correlations between different modalities (Hu et al. 2021; Chudasama et al. 2022; Yang et al. 2023) and ignore the importance of multi-granularity alignments (e.g., granularity such as frames, and relations between video frames or audio segments), which have been proven to be effective in other related multi-modal tasks, such as cross-modal retrieval (Li et al. 2021) and image-sentence matching (Xu et al. 2020; Liu et al. 2020). The hierarchical structures of both texts and other modalities advocate for weak correlation modeling. By exploring com-

*Corresponding Author.

positional semantics, helps to identify more vital but implicit correlations, e.g., correlations between an unaligned pair of entities and a group of video frames. On the other hand, multi-modal data inherently exhibits complex interrelations, with inconsistent semantic distributions (shown in Figure 1) among different modalities, thereby resulting in differences in sentiment congruity in both atomic-level and composition-level perspectives. Unfortunately, previous works overlooked the harm brought by such inconsistency for achieving better multimodal sentiment analysis.

To tackle these limitations, in this paper, we propose a novel **H**ierarchical **C**orrelation **M**odeling **Net**work (HCMNet) for multi-modal sentiment analysis. Specifically, our proposed method takes both atomic-level correlations between independent video frames, audio segments, and text tokens, as well as composition-level correlations considering spatial and semantic dependencies to explore weak dependency signals. To obtain atomic-level correlations, we design a dynamic attention reasoning method to align different modalities into the same space and compute the similarity score for each token-utterance-segment pair via inner products. Next, we obtain composition-level correlations based on the treated features of the text, audio, and video modalities acquired in the previous step. Concretely, we introduce a topological graph reasoning strategy, which constructs three uni-modal graphs using semantic dependencies among words and spatial dependencies among video utterances or audio segments to capture composition-level features for each modality using graph convolutional networks.

More importantly, we also mitigate the impact of distributional inconsistencies between modalities from both atomic-level and composition-level perspectives. From the atomic-level perspective, we design an atomic-level contrastive loss that empowers the model to learn robust class-relevant features in atomic-level feature space and alleviate the adverse effect of distributional inconsistency. As for composition-level, we propose a semantic optimal transport module to integrate transport flows with video, audio, and text graphs to constrain the composition-level incongruities between modalities. Specifically, we first utilize an optimal transport kernel to redefine the alignment problem across different modality pairs, eliminating the distributional gap between modalities by computing an informative cost matrix between video, audio, and text graphs. Then, we acquire optimal transportation plans, which are used for assigning source values to target distribution at minimum total cost. By doing so, it can learn strong cross-modal distributional consistency in composition-level features.

We validated our HCMNet on several benchmarks including CMU-MOSEI, IEMOCAP, and MELD over several models. Experiments demonstrate the effectiveness and universality of our approach, and extensive analyses provide insights into when and how our method works. In summary, the main contributions of this work are as follows:

- To the best of our knowledge, we are the first to exploit hierarchical semantic correlations between textual and visual modalities to jointly model the atomic-level and composition-level correlations for MSA task.
- We also mitigate the impact of distributional inconsistencies between modalities from both atomic-level and composition-level perspectives via contrastive learning and optimal transport learning.
- At the same time, the universality of our HCMNet method also provides the possibility to extend it to other multimodal understanding tasks.

Related Work

Multimodal Sentiment Analysis. Most MSA solutions adopt two different paradigms to understand multimodal emotion content. First, some of them paid more attention to designing advanced transformer architectures to capture the emotion dependencies across different modalities. UniMSE (Hu et al. 2022) proposed to obtain multimodal features that are fused by integrating audio and vision representations into a language model. MVN (Ma et al. 2022) proposed a multi-view network to explore both word-level and utterance-level emotion information. i-Code (Yang et al. 2023) designed an integrative and composable multimodal learning framework for triple-modal learning. MultiEMO (Shi and Huang 2023) integrated multimodal cues by capturing cross-modal mapping relationships. The second group is graph-based methods. MMGCN (Hu et al. 2021) leveraged both multimodal information and long-distance contexts for efficient emotion learning. AdaIGN (Tu et al. 2024) designed a graph interaction method to balance intra- and inter-speaker context dependencies for MSA task. However, most of them only consider atomic information to model the correlations contained in different modalities. Therefore, our work aims to emphasize the importance of more complex compositional information and the necessity of a combination of atomic-level and composition-level correlations.

Multimodal Fusion Methods. Early methods can be broadly categorized into two groups: aggregation-based methods (Hazirbas et al. 2017; Zeng et al. 2019; Valada, Mohan, and Burgard 2020; Colombo et al. 2021; Song et al. 2020) and methods that employ Optimal Transport (OT) (Chen et al. 2020; Pramanick, Roy, and Patel 2021; Zhou, Fang, and Feng 2023; Xu and Chen 2023). In the former, separate representations are learned for each modality, and these learned representations from different modalities are directly aggregated. However, these approaches lack effective inter-modal communication. These methods overlook the intra-modal characteristics by simply aligning distributions (Song et al. 2020). While some approaches (Ju et al. 2021; Han, Chen, and Poria 2021) attempted to combine both aggregation and alignment, they often require intricate hierarchical design, which can introduce additional computational costs and engineering complexity. OT-based works aim to achieve balanced feature alignment using the optimal transport method. CMOT (Zhou, Fang, and Feng 2023) conducted cross-modal mixup via optimal transport. OT-Coattn (Xu and Chen 2023) designed OT-based co-attention for structural interactions in survival prediction tasks. However, it is still a challenge to apply those methods for correlation learning. In this work, we propose a Graph-based optimal transport module to assist the model in learning semantic

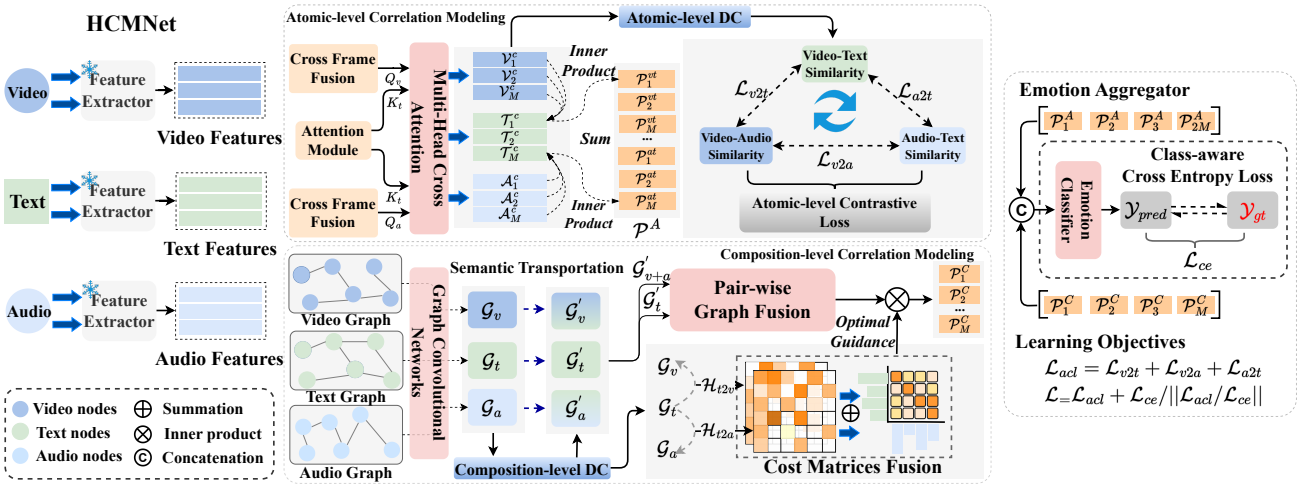

Figure 2: Illustration of the architecture of HCMNet. It consists of three components. (1) Feature Representation, which leverages existing feature extractors to provide raw multi-modal features. (2) Atomic-level Correlation Modeling (ACM) enhances the emotion semantics and performs distribution constraints at the atomic level. (3) Composition-level Correlation Modeling (CCM) aims to explore compositional congruity via graph learning and distribution constraints using optimal transport.

interactions and facilitate cross-modal communication.

Method

Preliminaries

Multimodal sentiment analysis aims to identify the specific emotion category to which a given sample (comprising video, text, and audio components) belongs. Formally given a multimodal feature vectors $\mathbf{X_i} = \{\mathbf{X_i^v}, \mathbf{X_i^t}, \mathbf{X_i^a}\}$, where v, t, a denote video, text, and audio modality, respectively, the goal of multimodal sentiment analysis is to predict the emotion label y_i^{gt} of the input feature vectors.

Feature Extraction

Given an input that contains three modalities $(\mathbf{X_i^v}, \mathbf{X_i^t}, \mathbf{X_i^a})$, we first employ the pre-trained RoBERTa (Liu et al. 2019) to produce a feature representation for each word token, denoted as $\mathbf{I_i^T} = [t_1, t_2, ..., t_n]$, where n is the number of word tokens and $\mathbf{I_i^T} \in \mathbb{R}^{n \times d_t}$. To learn contextual information, we follow previous work (Hu et al. 2022) to concatenate the current utterance with its former and latter 2-turn utterances. For video modality, we follow the previous insights (Shi and Huang 2023; Hu et al. 2022) and extract video visual features $\mathbf{I_i^V} \in \mathbb{R}^{T \times d_v}$ by employing the pre-trained efficientNet (Tan and Le 2019). Moreover, we feed treated features to additional Multilayer Perceptron (MLP) to model the frame importance and relationships contained in videos. As for the audio modality, we extract Mel-spectrogram sequential vectors by utilizing librosa toolkit [1] and fully-connected layers as audio features $\mathbf{I_i^A} \in \mathbb{R}^{k \times d_a}$.

[1] https://github.com/librosa/librosa

Atomic-level Correlation Modeling

To model the atomic-level correlations, we first encode each modality, mapping them into the same semantic space, and learn atomic-level correlation by dynamic attention reasoning method. After that, we introduce atomic-level contrastive learning to constrain the inconsistent distributions of each modality. The corresponding details are shown in the top branch of Figure 2.

For modality encoding, we employ the pre-trained video-text-audio CLIP (Guzhov et al. 2022) to encode the different unimodal features with modality-specific heads. Specifically, we use pre-trained CLIP (Radford et al. 2021) for visual and textual encoding, as well as pre-trained ESResNeXt (Guzhov et al. 2022) for audio encoding. Therefore, we can obtain the refined unimodal features. Moreover, we employ an attention Module for text features and a cross-frame Fusion Module for video and audio features to learn the long-term dependency in each modality.

$$\mathcal{T}^i = \mathbf{W_t}(\mathcal{F}^t_{direct}([\mathbf{I_i^T} \oplus \alpha_\mathbf{t}]) + \mathbf{b_t} \quad (1)$$

$$\mathcal{V}^i = \mathbf{W_v}(\mathcal{F}^v_{cross}([\mathbf{I_i^v} \oplus \beta_\mathbf{v}]) + \mathbf{b_v} \quad (2)$$

$$\mathcal{A}^i = \mathbf{W_a}(\mathcal{F}^a_{cross}([\mathbf{I_i^a} \oplus \beta_\mathbf{a}]) + \mathbf{b_a} \quad (3)$$

where W_* and b_* are the learnable matrices, α and β_* are the [CLS] vector. After learning the long-term dependencies, we can get the final video features $\mathcal{V}^i = \{\mathcal{V}^i_{(1)}, \mathcal{V}^i_{(2)}, ..., \mathcal{V}^i_{(M)}\}$, text features $\mathcal{T}^i = \{\mathcal{T}^i_{(1)}, \mathcal{T}^i_{(2)}, ..., \mathcal{T}^i_{(M)}\}$ and audio features $\mathcal{A}^i = \{\mathcal{A}^i_{(1)}, \mathcal{A}^i_{(2)}, ..., \mathcal{A}^i_{(M)}\}$, respectively.

To model the atomic-level correlation of $\mathbf{V_i}$, $\mathbf{T_i}$ and $\mathbf{A_i}$, we propose a dynamic attention reasoning process to apply cross-attention mechanisms to dynamically integrate different modalities, which are defined as:

$$\mathbf{h}_i^m = softmax(\frac{I_i^m \mathcal{T}_i^\top}{\sqrt{d_{ctn}}})\mathcal{T}_i + \lambda\alpha_t(\beta_v + \beta_a) \quad (4)$$

where I_i^m denotes input modality and $m \in \{v, a\}$. $h_i^m \in \{1, 2, ..., M_{ctn}\}$ is the i-th attention head, and M_{ctn} is hyper-parameter of the number of cross-attention network. λ is the CLS coefficient. Therefore, the atomic-level features are:

$$\begin{cases} \widetilde{\mathcal{T}_i^c} = \{\mathcal{T}_{(l)}^c\}_{l=1}^M \\ \widetilde{\mathcal{V}_i^c} = \mathcal{V}_i^c + Proj(\{\mathbf{h}_i^v\} \\ \widetilde{\mathcal{A}_i^c} = \mathcal{A}_i^c + Proj(\{\mathbf{h}_i^a\} \end{cases} \tag{5}$$

Then, the video and audio features are individually fused with text features using the element-wise inner product, yielding the final atomic-level representation $\mathcal{P}^A = [\mathcal{P}^{vt} \oplus \mathcal{P}^{at}] \in \mathbb{R}^{2M \times d_m}$, where $\mathcal{P}^{vt} = \{\mathcal{P}_{(1)}^{vt}, \mathcal{P}_{(2)}^{vt}, ..., \mathcal{P}_{(M)}^{vt}, \}$, $\mathcal{P}^{at} = \{\mathcal{P}_{(1)}^{at}, \mathcal{P}_{(2)}^{at}, ..., \mathcal{P}_{(M)}^{at}, \} \in \mathbb{R}^{M \times d_m}$.

$$\mathcal{P}^{m'} = \{\sum_{l=1}^{M} I_{(l)}^{m'} \mathcal{T}_{(i)}^c\}_{i=1}^M \tag{6}$$

where $m' \in \{vt, at\}$ and $I_i^{m'} \in \{\widetilde{\mathcal{V}_i^c}, \widetilde{\mathcal{A}_i^c}\}$, $\mathcal{P}^{m'}$ denotes the final atomic-level representations.

Atomic-level Distribution Constraint. Most previous works obtain multimodal features by directly fusing different unimodal representations. However, due to the semantic gaps among diverse modalities, a straightforward fusion approach may potentially elevate the vagueness of emotional information. Therefore, we propose to use contrastive learning for feature semantic distribution constraint, named Atomic-level Contrastive Learning (ACL). For the multimodal samples that have the same emotion class, we hope that they are in the same semantic space, and vice versa. Therefore, our ACL method can perform a unified embedding process with a contrastive learning manner, reducing the distributional inconsistency at the atomic level. Ultimately, given the encoded features $\widetilde{\mathcal{V}_i^c}$, $\widetilde{\mathcal{T}_i^c}$, and $\widetilde{\mathcal{A}_i^c}$, we can define the cosine similarity loss function ℓ_{ACL} between video, text, and audio representations as below:

$$\mathcal{L}_{acl} = \underbrace{\frac{<\widetilde{\mathcal{V}_i^c}, \widetilde{\mathcal{T}_i^c}>}{\| \widetilde{\mathcal{V}_i^c} \| \| \widetilde{\mathcal{T}_i^c} \|}}_{\mathcal{L}_{v2t}} + \underbrace{\frac{<\widetilde{\mathcal{V}_i^c}, \widetilde{\mathcal{A}_i^c}>}{\| \mathcal{V}_i^c \| \widetilde{\mathcal{A}_i^c} \|}}_{\mathcal{L}_{v2a}} + \underbrace{\frac{<\widetilde{\mathcal{A}_i^c}, \widetilde{\mathcal{T}_i^c}>}{\| \widetilde{\mathcal{A}_i^c} \| \| \widetilde{\mathcal{T}_i^c} \|}}_{\mathcal{L}_{a2t}} \tag{7}$$

Therefore, as is defined in $\mathcal{L}_{acl}$, for a multimodal pair, the small similarity means the features should be accordingly pushed away. Otherwise, they should be pulled close.

Composition-level Correlation Modeling

The composition-level correlation modeling considers the more complex structure of different modalities. To achieve that, we introduce topological graph reasoning, which utilizes topology graph structures in each modality to capture inter-modal correspondence most related to sentiment. Specifically, we first construct the topology graph structures for the textual, visual, and audio modalities. Then, we model the three different graphs with graph convolution network (GCN) (Kipf and Welling 2016). The details are shown in the bottom branch of Figure 2. Specifically, for the text graph $\mathcal{G}_t = (\mathcal{V}_\mathbf{t}, \delta_\mathbf{t})$, we consider tokens in the text features as graph nodes $\mathcal{V}_\mathbf{t}$ and employ dependency relations between words extracted by spaCy[2] as graph edges $\delta_\mathbf{t} \in \mathbb{R}^{M \times M}$, which have been emphasized to be resultful for various graph learning process. For constructing video graphs $\mathcal{G}_v = (\mathcal{V}_\mathbf{v}, \delta_\mathbf{v})$, we build edges between each utterance according to the cosine similarity of representations. As is defined in Equation 8, if the cosine similarity σ_{ij} between two utterances is greater than the threshold η, we establish an edge between the two utterances. For audio graph learning, we follow a simple utterance-to-node transformation, where M utterances that are short and overlapping segments of the audio signal are the M nodes in an audio graph $\mathcal{G}_a = (\mathcal{V}_\mathbf{a}, \delta_\mathbf{a})$. We simply utilize the line graph edge definition to construct the audio graph, which connects each utterance with the sequence order.

$$\delta_v = \begin{cases} \sigma_{ij} & , if\ \sigma_{ij} > \eta \\ 0 & , otherwise \end{cases} \tag{8}$$

Consequently, given a set of graphs ($\mathcal{G}_t$, $\mathcal{G}_v$, and $\mathcal{G}_a$), we resort to the graph convolution network (GCN) to mine the inherent relationship and learn the multi-modal composition-level semantics. We aim to learn the node features with their neighborhoods for fine-grained semantic mixing. The detailed formulas are defined below:

$$\begin{cases} \mathcal{V}_\mathbf{t}^{\mathbf{k}'} = \mathrm{Re}LU\left(\tilde{\delta}_\mathbf{t} \mathcal{V}_t^{k-1} \mathbf{W}_\mathbf{t}^\mathbf{k} + \mathbf{b}_\mathbf{t}^\mathbf{k}\right) \\ \mathcal{V}_\mathbf{v}^{\mathbf{k}'} = \mathrm{Re}LU\left(\tilde{\delta}_\mathbf{v} \mathcal{V}_v^{k-1} \mathbf{W}_\mathbf{v}^\mathbf{k} + \mathbf{b}_\mathbf{v}^\mathbf{k}\right) \\ \mathcal{V}_\mathbf{a}^{\mathbf{k}'} = \mathrm{Re}LU\left(\tilde{\delta}_\mathbf{a} \mathcal{V}_a^{k-1} \mathbf{W}_\mathbf{a}^\mathbf{k} + \mathbf{b}_\mathbf{a}^\mathbf{k}\right) \end{cases} \tag{9}$$

$$\tilde{\delta}_m = (\mathbf{D}_\mathbf{m})^{-\frac{1}{2}} \delta_m (\mathbf{D}_\mathbf{m})^{-\frac{1}{2}} \tag{10}$$

where $\tilde{\delta}_m$ is the normalized symmetric adjacency matrix, D_m is the degree matrix of adjacency matrix δ_m, $\mathcal{V}_t^{k'}$ is the k^{th} process of GCNs. $k \in [1, M_{gcn}]$, $m \in \{t, v, a\}$ denotes the different modalities. $\mathbf{W}_\mathbf{m}^\mathbf{l} \in \mathbb{R}^{d_h \times d_h}$ is the weight matrix and $\mathbf{b}_\mathbf{m}^\mathbf{l} \in \mathbb{R}^{d_h}$ is the bias matrix. l is the hyper-parameter of the number of GCN layers.

Composition-level Distribution Constraint. Next, we design the composition-level optimal transport module to learn the semantic distribution for video-text and audio-text groups, respectively. Different from the previous works (Xu and Chen 2023; Zhou, Fang, and Feng 2023), it is necessary to model the correlation between different pairs of modality. Therefore, we propose to learn multimodal descriptors by integrating video-audio graphs with text graphs under the guidance of transferred cost matrices obtained in the optimal transport process. Specifically, we first feed the final outputs of the GCN layers, $\mathcal{G}_m = \{g_m^{(1)}, g_m^{(2)}, ..., g_m^{(M)}\}$ into the optimal transport module. Formally, the optimal transport is conducted on audio-to-text and video-to-text and is defined by the discrete Kantorovich formulation to search the optimal semantic flows $\mathcal{H}_{m'}$ between graph $\mathcal{G}_v \in \mathbb{R}^{M \times d_m}$, $\mathcal{G}_a \in \mathbb{R}^{M \times d_m}$ and $\mathcal{G}_t \in \mathbb{R}^{M \times d_m}$:

$$\mathcal{W}(\mathcal{G}_{m'}, \mathcal{G}_t) = \min_{\mathcal{H}_{m'} \in \prod(\mu_{m'}, \mu_t)} < \mathcal{H}_{m'}, \mathcal{C}_{m'} > \tag{11}$$

[2]https://spacy.io/

$$\mathcal{G}^{'}_{m'} = \mathcal{G}_{m'} + \gamma_{m'}\mathcal{H}^{\top}_{m'}\mathcal{G}_{m'} \quad (12)$$

where $m' \in \{v, a\}$ and $\mathcal{C}_{m'}$ denote the cost matrices and are defined with Euclidean distance that measures the distance of local pair-wise instances of $\mathcal{G}_m$. $\mu_{m'}$ and μ_t are the marginal distributions. γ is the adaptive graph coefficient.

After obtaining the transport flows, the updated graphs can be defined as Equation 12. To reduce modality complexity and obtain semantic-consistent multimodal features, we feed the processed graph representations into a Pair-wise Graph Fusion module, an M_c-layer attention network, where M_c is the hyper-parameter. This allows us to learn the semantic dependencies between visual-acoustical and textual features. We concate visual graph $\mathcal{G}^{'}_v$ and audio graph $\mathcal{G}^{'}_a$ as query and key, and text graph as value for attention score calculation, which can be defined as the following equation.

$$\mathcal{G}_{va} = softmax(\frac{\mathbf{Q_{v+a}}\mathbf{K}^{\top}_{\mathbf{v+a}}}{\sqrt{d_{attn}}})\mathbf{V_t} \quad (13)$$

$$\mathcal{P}^u = \mathbf{W_1}(\mathbf{W_2}\mathcal{G}_{va} + \mathbf{b_2}) + \mathbf{b_1} \quad (14)$$

where $\mathbf{W_1}$, $\mathbf{W_2}$ are the weight parameters of feed-forward layers, and $\mathbf{b_1}$ and $\mathbf{b_2}$ are the bias parameters. $\mathbf{Q_{v+a}}, \mathbf{K_{v+a}} = \mathbf{W_q}[\mathcal{G}^{'}_v \oplus \mathcal{G}^{'}_a]$ and $\mathbf{V_t} = W_v\mathcal{G}_t$.

We fuse the visual-acoustical features $\mathcal{P}^u \in \mathbb{R}^{M\times d_m}$ with transferred cost matrix $\mathcal{C}' \in \mathbb{R}^{M\times M}$ to generate the final multimodal descripors $\mathcal{P}^C$. We conduct element-wise addition fusion to calculate the transferred cost matrix: $\mathcal{C}' = \mathcal{C}_v + \mathcal{C}_a$. The intention is to retrieve the crucial representations in inter- and intra-graph. The formula is as follows:

$$\mathcal{P}^C = \mathcal{C}' \odot \mathcal{P}^u \in \mathbb{R}^{M\times d_m} \quad (15)$$

where $\odot$ denotes vector inner-product and M is the number of visual-acoustical features, d_m is the dimension of $\mathcal{P}^u$.

Learning Objectives

To perform emotion analysis with dual-level features, we concatenate both the atomic- and composition-level features into the final feature representations for prediction. The Emotion Classifier contains a Class-aware Cross-Entropy Loss function $\mathcal{L}_{cce}$ between multimodal representations and emotional labels as is defined below:

$$\mathcal{L}_{cce} = -\sum \log \mathbf{P}_i \left(y = i|[\mathcal{P}^A \oplus \mathcal{P}^C]\right) + \xi||\theta||^2_2 \quad (16)$$

As for the learning objective, we introduce both the Atomic-level Contrastive Learning Loss $\mathcal{L}_{ACL}$ in Equation 7 and Class-aware Cross-Entropy Loss $\mathcal{L}_{cce}$ in Equation 16 for correlation learning. To balance the difference in the algebraic scale of the two losses, we adopt an adaptive loss formula. The detailed formula is as follows:

$$\mathcal{L} = \mathcal{L}_{acl} + \mathcal{L}_{cce} / \parallel \mathcal{L}_{acl}/\mathcal{L}_{cce} \parallel \quad (17)$$

where the $||\cdot||$ represents the truncated gradient operator, which calculates the adaptive balance coefficient of losses.

Dataset	Train	Valid	Test	All
MELD	9989	1108	2610	13707
IEMOCAP	5354	528	1650	7532
CMU-MOSEI	16326	1871	4659	22856

Table 1: The details of CMU-MOSEI, MELD, and IEMOCAP, including data splitting details.

Experimental Setup

Experimental Settings

Datasets. We assess the performance of our method on multimodal sentiment analysis benchmark datasets, including **IEMOCAP** (Busso et al. 2008) and **MELD** (Poria et al. 2018) and **CMU-MOSEI** (Zadeh et al. 2018) datasets. The statistics are reported in Table 1. Both of them are multimodal datasets with textual, visual, and acoustic modalities.

Implementation Details. For a fair comparison, we follow previous works (Hu et al. 2022; Ma et al. 2022) to pre-process the datasets and use the same dataset split. We employ ResNet-50 CLIP (Radford et al. 2021) for the visual and textual head, and ESResNeXt initialized on pre-trained datasets for the acoustic head. The number of training epochs is 100. We set the batch size as 64 for three datasets. We utilize AdamW as the optimizer with an initial learning rate of 2×10^{-4}. The dropout rate is set to 0.1 to avoid overfitting. The number of GCN layers is set to 4 as default. We use an 8-layer PGF module for graph fusion. All the experiments are carried out on NVIDIA RTX3090 GPUs. We use the weighted-F1 (w-F1) score as evaluation metrics for IEMOCAP and MELD datasets. For the CMU-MOSEI dataset, we adopt mean absolute error (MAE), Pearson correlation (Corr), accuracy (Acc), and F1-score as metrics. More details about datasets are provided in Appendix.

Baseline Models

To validate the effectiveness of HCMNet, we compared it with several state-of-the-art baselines. **DialogueRNN** (Majumder et al. 2019) and **DialogueGCN** (Ghosal et al. 2019) are dialogue-based models. They learn context information using recurrent networks and directed graphs. **ICCN** (Sun et al. 2020) learns correlations between three modalities via deep canonical correlation analysis. **IterativeERC** (Lu et al. 2020) enhances emotion interactions by using predicted emotion labels. **MMIM** (Han, Chen, and Poria 2021) hierarchically maximizes the Mutual Information in unimodal input pairs for the MSA task. **MMGCN** (Hu et al. 2021) leverages both multimodal information and long-distance contexts for efficient emotion learning. **UniMSE** (Hu et al. 2022) obtains multimodal features that are fused by integrating audio and vision representations into language models. **i-Code** (Yang et al. 2023) designs an integrative and composable multimodal learning framework for triple-modal learning. **MultiEMO** (Shi and Huang 2023) integrates multimodal cues by capturing cross-modal mapping relationships. **AdaIGN** (Tu et al. 2024) proposes a new adaptive graph learning for cross-modal interaction.

Methods	IEMOCAP							MELD							
	Happy	Sad	Neutral	Anger	Excited	Frustrated	w-F1	Neutral	Surprise	Fear	Sad	Joy	Disgust	Angry	w-F1
DialogueRNN	33.18	78.80	59.21	65.28	71.86	58.91	62.75	76.23	49.59	0.00	26.33	54.55	0.81	46.76	58.73
DialogueGCN	51.87	76.76	56.76	62.26	72.71	58.04	63.16	76.02	46.37	0.98	24.32	53.62	1.22	43.03	57.52
IterativeERC	53.17	77.19	61.31	61.45	69.23	60.92	64.37	77.52	53.65	3.31	23.62	56.63	19.38	48.88	60.72
MMGCN	42.34	78.67	61.73	69.00	74.33	62.32	66.22	-	-	-	-	-	-	-	58.65
UniMSE	-	-	-	-	-	-	70.66	-	-	-	-	-	-	-	65.51
MultiEMO	65.77	85.49	67.08	69.88	77.31	70.98	72.84	79.95	60.98	29.67	41.51	62.82	36.75	54.41	66.74
AdaIGN	53.04	81.47	71.26	65.87	76.34	67.79	70.74	-	-	-	-	-	-	-	-
HCMNet	**67.86**	**86.37**	**69.74**	**72.38**	**78.19**	**71.33**	**74.62**	**82.09**	**63.87**	**30.31**	**45.86**	**64.77**	**39.07**	**57.15**	**68.94**

Table 2: Results on IEMOCAP and MELD. The best and secondary performances are in bold and underlined, respectively.

Method	CMU-MOSEI				
	MAE ↓	Corr ↑	Acc-7 ↑	Acc-2 ↑	F1 ↑
ICCN	0.565	0.704	51.60	84.20	84.20
MMIM	0.526	0.772	54.24	85.97	85.94
UniMSE	0.523	0.773	54.39	87.50	87.46
i-Code	0.502	0.811	50.80	87.50	87.40
HCMNet	**0.489**	**0.819**	**54.97**	**87.63**	**87.56**

Table 3: Results on MOSEI dataset. The best and secondary performances are in bold and underlined, respectively.

Variant Model	IEMOCAP		MELD	
	ACC	w-F1	ACC	w-F1
HCMNet (ours)	73.86	74.62	68.31	68.94
ACM-only	70.13	70.25	67.97	68.36
CCM-only	68.55	68.76	67.24	67.31
w/o ACL	69.71	70.22	67.08	67.45
w/o graph learning	69.82	70.18	67.33	67.74
w/o CDC	68.83	69.52	66.67	67.49

Table 4: Experiment results of ablation study. We compare different variants on IEMOCAP and MELD datasets.

Results and Analysis

Main Results

We compare our HCMNet model with the existing methods. To verify the capability of multi-granularity multimodal sentiment representation, we design experiments carried out on IEMOCAP and MELD datasets that have fine-grained sentiment categories for multimodal sentiment analysis. Further, we conduct both 7-class and 2-class sentiment analysis experiments on the CMU-MOSEI dataset. Table 2 and Table 3 show the comparison results on the three datasets. We can obtain the following conclusions: **(1)** Our method achieves the best performance across three datasets, demonstrating the effectiveness of learning hierarchical correlations with distribution constraints for MSA task. Specifically, HCMNet obtains 2.44% and 3.30% overall Weighted-F1 improvement compared with MultiEmo (Shi and Huang 2023) in IEMOCAP and MELD datasets, respectively. **(2)** Table 3 shows the detailed performance comparison on the CMU-MOSEI dataset. We note that our method achieves mediocre accuracy and F1-score improvement, which indicates that there is potential space for HCMNet to optimize the emotion representation process for the CMU-MOSEI dataset. **(3)** The F1 score of MELD dataset is significantly lower than the F1-score on IEMOCAP dataset. The possible reason is that IEMOCAP is collected from well-designed utterances that have selected scripts with clear emotional content.

Ablation Study

To investigate the effectiveness of the different components and settings, we introduce several variants of our method for comparison on IEMOCAP and MELD datasets.

(1) Analysis of model components. We compared HCMNet with the following derivations. **ACM-only** contains a single atomic-level correlation modeling module. **CCM-only**, which mainly utilizes graph learning and composition-level optimal transport to obtain multimodal features. **w/o ACL** means we remove the Atomic-level Contrastive Learning from the ACM module, which aims to explore the impact of atomic-level distribution constraint. **w/o graph learning**: we replace graph learning with untreated features gained from the ACM module directly. To testify to the effectiveness of optimal transport in composition-level correlation modeling, we remove the composition-level distribution constraint, which is **w/o CDC**. The results are shown in Table 4 and we make the following observations: **First**, compared with the variant that only uses the CCM module, ACM-only model which utilizes atomic-level correlation modeling and distribution constraint for MSA task gains more positive results on both IEMOCAP and MELD datasets. Therefore, fine-grained contrastive learning and element-wise fusion help a lot when learning the clarity correlations. **Second**, Atomic-level Contrastive Learning (ACL) plays an important role in catching the semantics that are exploited for particle semantic aligning in subsequent features representation process. **Third**, it is more befitting to employ an efficient graph learning approach to obtain structural semantic information. We can infer that composition-level distribution constraint balances different semantic distributions between different modality pairs.

(2) Analysis of different modality settings. We conduct an ablation experiment with four modality settings on IEMOCAP and MELD datasets, which gives us more insights and inspiration about the MSA task and the attribute of the dataset. Table 5 shows the experiment result. We can infer the following conclusions. **First**, performances on both two datasets vary with different modality settings. As shown in

Modality	IEMOCAP		MELD	
	ACC	w-F1	ACC	w-F1
A+T	71.20	71.71	65.09	65.12
V+T	72.61	73.22	66.98	67.19
V+A	71.96	72.07	65.42	66.48
V+A+T	**73.86**	**74.62**	**68.31**	**68.94**

Table 5: Ablation study of HCMNet with different modalities on IEMOCAP and MELD datasets. T, V and A represent textual, visual and acoustic modalities, respectively.

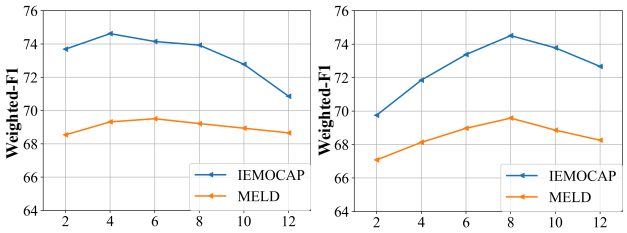

Figure 3: Parameter sensitive analysis of the number of GCN (left) and PGF layers (right).

Table 5, experiment items with textual modality generally gain satisfactory scores, indicating that dialogue content is more important in multimodal emotion prediction. Specifically, compared with **V+A**, the combination of **V+A+T** achieves 2.64% and 4.03% weighted-F1 increase on the two datasets, respectively. **Moreover**, for two-modality settings, it can be observed that **V+T** is more promising in emotion semantic modeling compared to others.

Parameter Sensitivity Analysis

We evaluate the effect of the number of GCN and PGF layers. As is shown in Figure 3. We can observe that as the number of GCN and PGF layers increases, the model's performance improves steadily until reaching its peak, after which it starts to decline. Therefore, we can infer that: **(1)** Our model works best when the layers parameter is set to 4 for GCN and 8 for the PGF module. **(2)** It is ideal to choose fewer GCN layers and more PGF layers for better performance, indicating that graph learning used at the top of the module is less demanding than semantic-level graph fusion.

Qualitative Analysis

Visualization analysis. In Figure 4, we visualize the latent features obtained from HCMNet to explore how it works for MSA task. **(1)** The results reveal that the HCMNet is capable of capturing discriminative features for different categories. The distinct feature clusters for each category indicate that our model learns to highlight and differentiate between specific multimodal features associated with different emotion semantics. **(2)** Classes with CDC tend to exhibit closer proximity in the latent space, reflecting the model's ability to align multimodal features with emotional semantic relationships. As shown in Figure 4 (a), many outlier features spread across different feature spaces, thus disturbing the discriminability of emotion analysis models. When we employ the CDC module in HCMNet, as shown in Figure 4 (b), different emotion features maintain sufficient distances from each other, which provides beneficial partition boundaries.

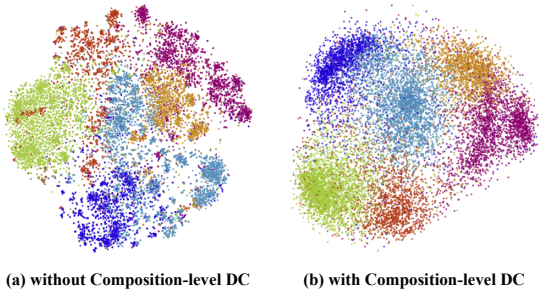

Figure 4: t-SNE visualization of IEMOCAP dataset. Different colored dots represent samples with different categories.

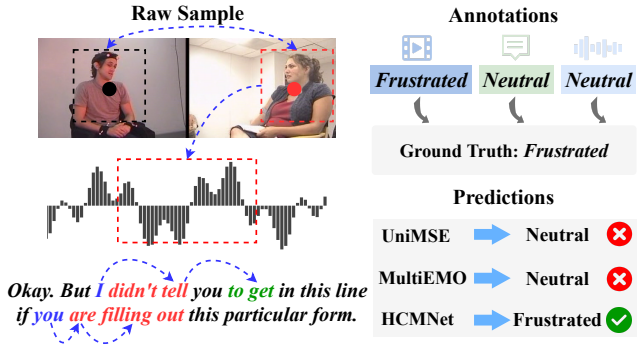

Figure 5: Case study of IEMOCAP dataset. We provide a sample to demonstrate the effectiveness of our method.

Case Study. To further demonstrate the effectiveness of our method, we visualize an example of MSA results on the IEMOCAP dataset in Figure 5. Note that the connections in the figure indicates the composition-level correlations of each modality. From the case, we can see two people talking face to face, with confused and frustrated emotions. Compared with previous work UniMSE (Hu et al. 2022) and MultiEMO (Shi and Huang 2023), our method considers both atomic- and composition-level information, thus giving accurate prediction. With hierarchical correlation modeling and semantic distribution constraints, our model can obtain more discriminative features for correlation modeling.

Conclusion And Future Work

We present a Hierarchical Correlation Modeling Network, enhancing the multimodal sentiment analysis by exploring both atomic- and composition-level correlations. Besides, we propose to perform distribution constraints for dual-level features. Experiments on three public benchmark datasets have demonstrated the superiority of the HCMNet model over previous methods. In future work, we will incorporate contextual knowledge from large vision-language models, thereby better learning emotional correlations.

Acknowledgments

This work is partially supported by the grants from the Natural Science Foundation of Shandong Province (ZR2024MF145, ZR2024ZD20), and the National Natural Science Foundation of China (62072469).

References

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42: 335–359.

Chen, L.; Gan, Z.; Cheng, Y.; Li, L.; Carin, L.; and Liu, J. 2020. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, 1542–1553. PMLR.

Chudasama, V.; Kar, P.; Gudmalwar, A.; Shah, N.; Wasnik, P.; and Onoe, N. 2022. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4652–4661.

Colombo, P.; Chapuis, E.; Labeau, M.; and Clavel, C. 2021. Improving multimodal fusion via mutual dependency maximisation. *arXiv preprint arXiv:2109.00922*.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

Guzhov, A.; Raue, F.; Hees, J.; and Dengel, A. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 976–980. IEEE.

Han, W.; Chen, H.; and Poria, S. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*.

Hazirbas, C.; Ma, L.; Domokos, C.; and Cremers, D. 2017. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I 13*, 213–228. Springer.

Hu, G.; Lin, T.-E.; Zhao, Y.; Lu, G.; Wu, Y.; and Li, Y. 2022. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256*.

Hu, J.; Liu, Y.; Zhao, J.; and Jin, Q. 2021. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. *arXiv preprint arXiv:2107.06779*.

Ju, X.; Zhang, D.; Xiao, R.; Li, J.; Li, S.; Zhang, M.; and Zhou, G. 2021. Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 4395–4405.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Li, Y.; Zhou, H.; Yin, Y.; and Gao, J. 2021. Multi-label pattern image retrieval via attention mechanism driven graph convolutional network. In *Proceedings of the 29th ACM international conference on multimedia*, 300–308.

Liu, C.; Mao, Z.; Zhang, T.; Xie, H.; Wang, B.; and Zhang, Y. 2020. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10921–10930.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lu, X.; Zhao, Y.; Wu, Y.; Tian, Y.; Chen, H.; and Qin, B. 2020. An iterative emotion interaction network for emotion recognition in conversations. In *Proceedings of the 28th international conference on computational linguistics*, 4078–4088.

Ma, H.; Wang, J.; Lin, H.; Pan, X.; Zhang, Y.; and Yang, Z. 2022. A multi-view network for real-time emotion recognition in conversations. *Knowledge-Based Systems*, 236: 107751.

Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; and Cambria, E. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 6818–6825.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Pramanick, S.; Roy, A. B.; and Patel, V. M. 2021. Multimodal learning using optimal transport for sarcasm and humor detection. 2022 IEEE. In *CVF Winter Conference on Applications of Computer Vision (WACV)(2021)*, 546–556.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Shi, T.; and Huang, S.-L. 2023. MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14752–14766.

Song, S.; Liu, J.; Li, Y.; and Guo, Z. 2020. Modality compensation network: Cross-modal adaptation for action recognition. *IEEE Transactions on Image Processing*, 29: 3957–3969.

Song, T.; Zheng, W.; Liu, S.; Zong, Y.; Cui, Z.; and Li, Y. 2021. Graph-embedded convolutional neural network for image-based EEG emotion recognition. *IEEE Transactions on Emerging Topics in Computing*, 10(3): 1399–1413.

Sun, Z.; Sarma, P.; Sethares, W.; and Liang, Y. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8992–8999.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Tu, G.; Xie, T.; Liang, B.; Wang, H.; and Xu, R. 2024. Adaptive Graph Learning for Multimodal Conversational Emotion Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19089–19097.

Valada, A.; Mohan, R.; and Burgard, W. 2020. Self-supervised model adaptation for multimodal semantic segmentation. *International Journal of Computer Vision*, 128(5): 1239–1285.

Veltmeijer, E. A.; Gerritsen, C.; and Hindriks, K. V. 2021. Automatic emotion recognition for groups: a review. *IEEE Transactions on Affective Computing*, 14(1): 89–107.

Xu, X.; Wang, T.; Yang, Y.; Zuo, L.; Shen, F.; and Shen, H. T. 2020. Cross-modal attention with semantic consistence for image–text matching. *IEEE transactions on neural networks and learning systems*, 31(12): 5412–5425.

Xu, Y.; and Chen, H. 2023. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21241–21251.

Yang, Z.; Fang, Y.; Zhu, C.; Pryzant, R.; Chen, D.; Shi, Y.; Xu, Y.; Qian, Y.; Gao, M.; Chen, Y.-L.; et al. 2023. i-code: An integrative and composable multimodal learning framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10880–10890.

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 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 (Volume 1: Long Papers)*, 2236–2246.

Zeng, J.; Tong, Y.; Huang, Y.; Yan, Q.; Sun, W.; Chen, J.; and Wang, Y. 2019. Deep surface normal estimation with hierarchical RGB-D fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6153–6162.

Zhang, W.; He, X.; and Lu, W. 2019. Exploring discriminative representations for image emotion recognition with CNNs. *IEEE Transactions on Multimedia*, 22(2): 515–523.

Zhou, Y.; Fang, Q.; and Feng, Y. 2023. CMOT: Cross-modal mixup via optimal transport for speech translation. *arXiv preprint arXiv:2305.14635*.