Multi-Task Momentum Distillation for Multimodal Sentiment Analysis

Ronghao Lin and Haifeng Hu

***Abstract*—In the field of Multimodal Sentiment Analysis (MSA), the prevailing methods are devoted to developing intricate network architectures to capture the intra- and inter-modal dynamics, which necessitates numerous parameters and poses more difficulties in terms of interpretability in multimodal modeling. Besides, the heterogeneous nature of multiple modalities (text, audio, and vision) introduces significant modality gaps, thereby making multimodal representation learning an ongoing challenge. To address the aforementioned issues, by considering the learning process of modalities as multiple subtasks, we propose a novel approach named Multi-Task Momentum Distillation (MTMD) which succeeds in reducing the gap among different modalities. Specifically, according to the abundance of semantic information, we treat the subtasks of textual and multimodal representations as the teacher networks while the subtasks of acoustic and visual representations as the student ones to present knowledge distillation, which transfers the sentiment-related knowledge guided by the regression and classification subtasks. Additionally, we adopt unimodal momentum models to explore modality-specific knowledge deeply and employ adaptive momentum fusion factors to learn a robust multimodal representation. Furthermore, we provide a theoretical perspective of mutual information maximization by interpreting MTMD as generating sentiment-related views in various ways. Extensive experiments illustrate the superiority of our approach compared with the state-of-the-art methods in MSA.**

***Index Terms*—Multi-task learning, knowledge distillation, unimodal momentum model, multimodal sentiment analysis.**

I. Introduction

MULTIMODAL Sentiment Analysis (MSA) has emerged as a pivotal and challenging area of research in recent years [1], [2], [3], [4]. With the rapid growth of user-generated online content on social media platforms, MSA has attracted increasing attention in multimodal applications encompassing textual (language), acoustic (human voice), and visual (facial expressions) modalities. In essence, the utilization of multiple modalities with complementarity for analyzing the same data object offers distinctive signals to facilitate semantic and emotional disambiguation [5], [6]. By simultaneously incorporating different modalities in the network, MSA demonstrates enhanced accuracy in inferring the sentiment intensity compared to text-only sentiment analysis which relies on a solitary and insufficient modality [7].

Owing to the wealth of semantic context embedded in language, the textual modality assumes a critical role in the field of MSA [8], [9]. Besides, recent advancements have showcased remarkable performance in downstream natural language processing tasks thanks to the large pre-trained Transformer-based language models such as BERT [10]. As a result, the textual representation is widely acknowledged as the most crucial representation [3], [11] in contrast to the representation of other modalities, which rely on manual feature extraction methods [12]. In other words, the textual modality is considered as the dominant modality, whereas the acoustic and visual modalities are deemed as inferior modalities in MSA. However, as mentioned by Zadeh et al. [1] and Wang et al. [8], exclusive reliance on the textual modality leads to subjective and biased emotion problems. Thus, the supplement of acoustic and visual modalities becomes imperative to encode multimodal representations and predict the sentiment accurately. During this process, the heterogeneous nature and the imbalanced information entropy of modalities result in significant gaps in multimodal representation modeling [3], which illustrates the disparities among the representations of different modalities within a shared distributional space. The primary challenge in MSA lies in bridging the modality gap and further exploring intra-modal dynamics (modality-specific features) and inter-modal dynamics (modality-shared features) [1].

To address the challenge posed by the modality gap, previous methods can be divided into forward guidance and backward guidance methods according to the modeling techniques for the exploration of intra- and inter-modal dynamics in representation learning. Forward guidance methods [1], [13], [14], [15], [16] concentrate on designing sophisticated modules to formulate the relationship between unimodal and multimodal representations. For instance, Tsai et al. [13] utilize modality-specific generative factors and multimodal discriminative factors to factorize representations while Han et al. [16] hierarchically maximize the mutual information at the level of unimodal input and multimodal fusion. Despite these advancements, these modeling strategies lack modality-specific supervision due to the sole multimodal annotation, thus limiting the extraction of intra-modal dynamics. Moreover, the intricate architectures of these networks require a large number of parameters and computational costs during the inference stage, imposing a considerable burden in practical applications [17]. Besides, the complex modules in these networks hinder the interpretability of multimodal representation

Manuscript received 4 December 2022; revised 23 May 2023; accepted 30 May 2023. Date of publication 2 June 2023; date of current version 23 May 2024. This work was supported by the National Natural Science Foundation of China under Grant 62076262. *(Corresponding author: Haifeng Hu.)* Recommended for acceptance by E. Cambria.

The authors are with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China (e-mail: linrh7@mail2.sysu.edu.cn; huhaif@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/TAFFC.2023.3282410

learning [18]. In contrast, backward guidance methods [3], [19] present additional prior constraints such as loss functions or subspace segmentation to capture intra- and inter-modal dynamics for multimodal representations. Specifically, Hazarika et al. [3] project multiple modalities into modality-specific and -invariant subspace, whereas Yu et al. [19] generate unimodal annotations to lead the unimodal representation learning in a self-supervision way. However, the explicit separation of subspaces proposed by [3] may struggle to effectively represent the modality-specific difference, potentially leading to confusion in the interaction among the modalities [19]. In addition, self-computed unimodal labels [19] need a hand-craft pseudo-label generating strategy and only offer hard supervision at the prediction level for the learning process of unimodal representations, which restricts the exploration of modality-specific information due to lacking supervision at the representation level.

Inspired by the concept of backward guidance, we propose a novel approach *Multi-Task Momentum Distillation (MTMD)* aiming at extracting modality-specific features through momentum distillation and exploring the interaction among multiple modalities by loss functions. Diverse from previous backward guidance methods, we regard the learning process of multiple modalities as distinct unimodal subtasks and the modeling of multimodal representations as the multimodal subtask. Furthermore, we adopt knowledge distillation in this multi-task framework to transfer modality-shared information among multiple subtasks without generating explicit hard unimodal labels. Notably, our approach achieves these goals without the need to increase computational parameters in inference. Moreover, the network architecture of modeling unimodal representations is hardly modified due to the sight of subtasks, which indicates that our approach can be effectively employed in various downstream tasks in the field of multimodal machine learning [20].

For the knowledge distillation of the multi-task framework, due to the plentiful semantic information contained in textual and multimodal representations, we deem the learning process of textual and multimodal representations as the teacher networks, while the modeling of acoustic and visual representations is considered as the student networks. To distill knowledge from the teacher networks to the student networks, we present *Dominant Modality Distillation (DMD)* and *Multimodal Fusion Distillation (MFD)*. Specifically, we utilize textual and multimodal subtasks to instruct the learning of acoustic and visual subtasks at both prediction-level and feature-level, which promotes the inferior modalities to concentrate on sentiment-related information. DMD involves response-based and feature-based knowledge distillation, in which the former is devoted to learning the sentiment prediction distribution through regression and classification heads, while the latter aims at encoding the representations using contrastive learning based on the sentiment classes. By doing so, we enable the transfer of the consistent and complementary modality-shared information between the dominant modality and inferior modalities. However, due to the subjective and biased emotion issues of textual modality, the guidance provided by textual subtask may potentially mislead the acoustic and visual subtasks. To address the issue, we introduce unimodal momentum models as exponential-moving-average counterparts of unimodal online models for acoustic and visual subtasks. The momentum models generate pseudo-targets, including unimodal features and prediction labels, to encourage the inferior modalities to retain the modality-specific features. In MFD, we assign learnable weight factors to the unimodal representations in multimodal fusion, which allow for adaptive adjustment of the contribution of different modalities. In addition, we leverage multimodal subtask as another teacher network to guide the unimodal subtasks, incorporating a hard samples attention strategy to enhance convergence. Nevertheless, the multimodal subtask encounters an unstable inference problem during the initial stages of training due to the random initialization of interaction layers. To mitigate this instability issue, we update the learnable weight factors in a momentum manner, as the momentum-updated strategy employed in the unimodal momentum model. In conclusion, our approach obtains a more robust multimodal representation and improves performance on sentiment prediction tasks compared to the state-of-the-art methods. Needless of modifying the network architecture for modeling unimodal inputs, the proposed MTMD retains applicability and generalizability in other multimodal machine learning tasks.

Briefly, the novel contributions of our approach can be summarized as:

- *A novel multi-task momentum distillation approach for MSA:* By regarding the learning of unimodal and multimodal representations as distinct subtasks, we propose a novel approach named MTMD to transfer the sentiment-related knowledge among the modalities, which successfully captures inter-modality dynamics and explores modality-shared information.
  1) Dominant Modality Distillation (DMD) utilizes the semantic knowledge of the textual subtask to guide the learning process of the inferior subtasks through response-based and feature-based knowledge distillation, which competently extracts the cross-modal features from different modalities.
  2) Multimodal Fusion Distillation (MFD) adaptively adjusts the contribution of multiple modalities using learnable weight factors, which enables efficient interaction of multimodal representations with the sentiment-related features during the stage of multimodal fusion.
  3) With the guidance of momentum distillation, the unimodal momentum models promote DMD to extract modality-specific information and encourage MFD to jointly learn robust multimodal representations for downstream sentiment prediction.
- *Interpretability based on mutual information maximization perspective:* From a mutual information maximization perspective, the presented ideas of multi-task learning and knowledge momentum distillation are interpreted as generating multiple sentiment-related views, illustrating the generalizability of the proposed MTMD in other multimodal learning tasks.
- *Surpassing current state-of-the-art methods:* Extensive experiments conducted on public datasets demonstrate that MTMD outperforms existing state-of-the-art methods.

II. Related Work

A. Multimodal Sentiment Analysis

Recently in the field of Natural Language Processing (NLP), to improve interpretability of deep neural networks, neurosymbolic AI framework [21], [22] has been proposed to conduct explainable natural language understanding, especially for sentiment analysis as a conventional language understanding task [23]. Meanwhile, with the surge of multimodal data, attempts to process multimodal signals with symbolic knowledge integrated model has attracted gradual attention in the NLP community [24], [25], increasing necessary demands for tackling Multimodal Sentiment Analysis (MSA) task in the research of sentiment analysis and emotion recognition.

Concretely, the MSA task integrates text, audio, and vision information to comprehend and predict the human sentiment (i.e., positive or negative)) contained in the multimodal data [20], [26], [27]. Previous surveys emphasize multimodal fusion as the core component of MSA and propose various fusion methods to learn the inter-modal dynamics and predict the accurate sentiment [4]. The fusion strategies can be classified into early fusion [1], [28], [29], [30], [31] and late fusion [19], [32], [33], [34], [35], [36]. Early fusion manipulates the input features from different modalities and interacts modality-shared information at the feature-level before multimodal concatenation. While due to the inconsistency of diverse modality space, early fusion suffers from the overfitting problem at the beginning of fusion. To avoid this issue, late fusion performs integration by decision voting after each modality outputs the prediction independently. However, late fusion neglects the low-level interaction among the modalities and lacks effective exploration of inter-modal dynamics. To promote strengths and avoid weaknesses of early fusion and late fusion, increasingly recent researches concentrate on the methods of hybrid fusion [2], [9], [16], [37], [38], [39], which adopts a hierarchical architecture to the multimodal network and presents fusion from input-level to output-level. Nevertheless, hybrid fusion depends on the sophisticated design of the fusion network, which tends to require higher computation costs and more parameters to be adjusted. Furthermore, the complexity of the network brings more difficulty in the interpretability of the learning process and the contribution of each modality.

To capture the inter-modal dynamics, we propose a novel approach named MTMD by interacting the modality information at both prediction-level and feature-level. Different from previous hybrid fusion methods, we take almost no change to the network architecture and focus on the interaction of multiple subtasks by regarding the learning process of different modalities as different subtasks. Under the guidance of multi-task learning and knowledge transferring, we successfully extract the modality-shared and sentiment-related features of the multimodal data. Moreover, for the purpose of exploring the modality-specific features, we introduce the momentum unimodal models as the moving-average counterparts of the unimodal online models, which retain the unimodal information without increasing neurons in the networks. With the effectiveness of achieving the state-of-the-art performance in two public MSA datasets, our approach brings new sights to recent commensense-based neurosymbolic AI framework [23] on sentiment analysis and emotion recognition with multimodal data.

B. Multi-Task Learning

Multi-task learning is a machine learning paradigm aiming at leveraging informative knowledge contained in different but related subtasks to improve model's generalization performance on multiple tasks [40]. Historically, the technologies of deep multi-task learning are divided into hard and soft parameter-sharing schemes according to the design of feature-sharing mechanisms in multi-task networks [41]. In hard parameter-sharing [42], [43], the networks consist of shared feature encoders and multiple task-specific heads to the corresponding tasks. In soft parameter-sharing [44], [45] on the other hand, the networks assign task-specific sets of parameters and handle cross-task constraints by regularization techniques. Recently, multi-task learning is broadly implemented in multimodal learning research [46], [47], [48]. Inspired by the previous works, we adopt the hard parameter-sharing strategy for the shared encoders for unimodal and multimodal subtasks and utilize two task-specific heads for sentiment regression and classification goals.

Compared with single-task learning, the primary challenge of multi-task learning is to balance the joint training process of different tasks [49], [50]. Since we consider the representation learning of different modalities as sentiment-related subtasks, inspired by the concept of task balancing, we adaptively adjust the contribution of different modalities by inserting learnable weight factors into multimodal fusion. Through multi-task supervisory signals, our approach learns a robust multimodal representation that efficiently extracts the modality-shared information.

C. Knowledge Distillation

Knowledge distillation transfers the knowledge from one deep neural network (the teacher) into another (the student) by minimizing the distance of the teacher's and student's output distribution [51]. As stated in Gou et al. [52], the forms of knowledge can be categorized into response-based [53], feature-based [54], and relation-based knowledge [55]. The former two are related to the outputs of specific layers in the teacher network while the last one refers to the relationships between different layers or data samples [52]. In particular, the response-based knowledge distillation performs distillation on the logits of the last output layer, which are also named neural responses. Considering the complex interdependencies of the structural representation dimensions, the feature-based knowledge distillation concentrates on the feature representations output from multiple layers of the network to capture higher-order correlations [56].

In the multimodal scenarios, modality-shared knowledge distillation further extends the knowledge transferring on different modalities [57], [58]. Besides, ALBEF [59] proposes momentum distillation in a pre-trained visual language model to extend knowledge distillation with a momentum model to prevent the online model from overfitting on the noisy web data. Aiming

at image-text matching, ALBEF deems image and text as equal for unimodal encoders and utilizes a cross-attention mechanism for the multimodal encoder to learn the modality relation jointly as a forward guidance method. The momentum distillation designed by ALBEF focuses only on two objective learning losses separately for unimodal encoders and multimodal encoder as independent single tasks, lacking the assistance of multi-task learning to interact the features of different modalities or the knowledge distillation among different unimodal and multimodal representations.

Differently, in this article, due to the rich sentiment semantic knowledge contained in textual modality and multimodal feature, we regard the textual subtask and multimodal subtask as the teacher network while the acoustic and visual subtask as the student network. To efficiently transfer the information from teacher to student, we conduct both response-based and feature-based knowledge distillation on the multiple subtasks across different modalities as a backward guidance method, needless of other parameters to explore inter-modal dynamics. In addition, with the great performance of contrastive learning on representational extraction [17], [56], [60], we leverage knowledge contrastive distillation on the representations of different modalities according to corresponding sentiment classes. Furthermore, inspired by but diverse from ALBEF which faces the problem of explicit noisy data, we present momentum distillation by designing the momentum unimodal model for the inferior modalities to retain the intra-modal dynamics under implicit noisy supervision of the dominant modality. The knowledge distillation strategy greatly reduces the modality gap, leading to effective extraction and interaction of sentiment-related information across multiple modalities.

III. Approach

The proposed approach is described in detail in this section. We first define the specific multimodal task and explain the corresponding notations of the problem. Then we present the whole multi-task architecture of the proposed MTMD, followed by the setup of unimodal subtasks and the the notations of unimodal momentum model. In addition, two primary modules of MTMD are introduced concretely. Finally, we delineate the total optimization objectives.

A. Task Definition

Multimodal Sentiment Analysis (MSA) aims at utilizing multimodal signals to predict sentiment intensity in the form of an emotion score. The input of the MSA task is utterances $X_u \in \mathbb{R}^{\ell_u \times d_u}$ derived from raw video fragments, where ℓ_u denotes the sequence length and d_u denotes the representation dimension of modality u. Specifically, $u \in \{t, a, v\}$ represents three types of signals including textual, acoustic, and visual modalities. Then the proposed approach integrates the sentiment-related information of these modalities into multimodal representation, and finally outputs the predicted sentiment scores $\hat{y}$ to reflect the accurate affective strength, where the ground truth sentiment label is denoted as y.

B. Multi-Task Momentum Distillation Architecture

As shown in Fig. 1, we construct the architecture of the proposed Multi-Task Momentum Distillation (MTMD) by two key ideas, which are *Multi-Task Learning* and *Knowledge Momentum Distillation*. The former aims at building multiple tasks to train the model to be capable of dealing with different forms of data and learn a more robust multimodal representation for sentiment-related generalization. Meanwhile, the latter focuses on transferring task-related knowledge from the teacher network to the student network in a momentum-updated way, making the transfer of knowledge more smoothly and efficiently.

Since previous research has demonstrated the effectiveness of multi-task learning methods [19], we train the proposed MTMD by multiple subtasks designed in the levels of multi-modalities and prediction heads.

- On the one hand, the sentiment inference process of different modalities can be regarded as multiple subtasks each of which outputs the sentiment labels predicted by unimodal representation. Similar to *unimodal subtasks*, the sentiment prediction from multimodal representation is seen as *multimodal subtask*, which represents the final prediction result in both validation and testing stages. The usage of these outputs from different subtasks becomes the key issue in balancing the contribution of each modality and learning a more robust unimodal representation.
- On the other hand, the problem of value regression can be converted into a classification task [61], meaning that the regression of sentiment scores can be deemed as the classification of sentiment class. Therefore, we utilize *regression and classification heads* to predict the sentiment intensity of the subtask of each modality.

To further transfer the knowledge among different subtasks, we introduce knowledge momentum distillation as a bridge to link different subtasks and cross the modality gap. Considering the predominant role of textual modality and fused multimodal representation, we account them as teacher networks and present two modules for knowledge distillation, i.e., dominant modality distillation and multimodal fusion distillation.

- Due to the rich semantic nature and powerful pre-trained language models as textual encoders, the textual modality is observed as the dominant modality in the MSA task which is indicated in previous works [3], [9]. We apply *Dominant Modality Distillation (DMD)* to instruct the encoding of acoustic and visual modality to learn in a similar way as the semantic representation of textual modality. Besides, the distillation of high-level semantic information promotes the inferior modalities encoder to concentrate on task-related features and filter the task-unrelated noise in the meanwhile.
- Additionally, the multimodal subtask fuses information from three modalities into the multimodal representation for the final sentiment prediction. As a more instructive subtask, the multimodal representation and prediction are seen as another teacher network similar to the textual subtask to guide the acoustic and visual subtasks to explore more commonalities among different modalities.

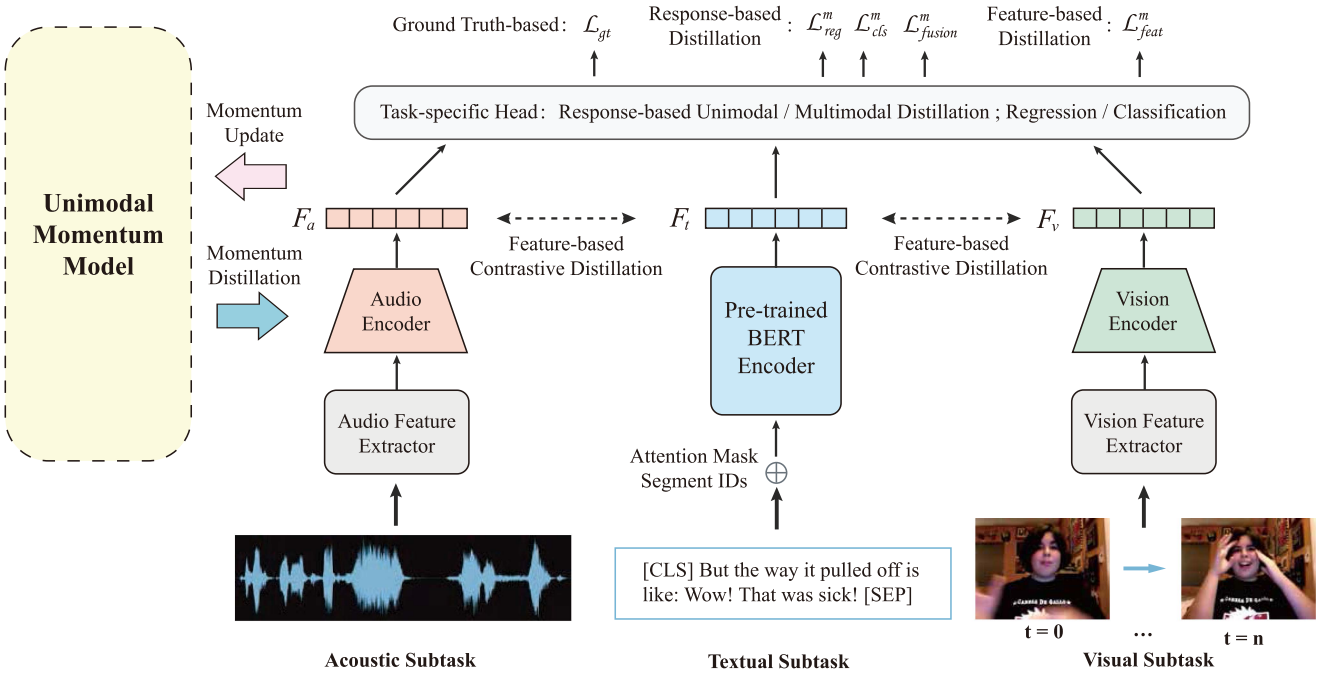


Fig. 1. Overall architecture of the proposed MTMD framework. Inputting utterances including acoustic, textual, and visual modalities, the online models for each unimodal subtask output the final unimodal representations F_u where $u \in \{t, a, v\}$. Then the feature-based contrastive distillation conducts contrastive learning between F_t and F_a/F_v which is optimized by $\mathcal{L}^m_{feat}$. Next, the unimodal representations from different modalities are weighted concatenated, and mapped into multimodal representation F_f in the multimodal subtasks. Lastly, both F_u and F_f are sent into task-specific heads for the final predictions of regression and classification. At the level of predictions, the response-based unimodal and multimodal distillation are conducted and optimized by $\mathcal{L}^m_{reg}$/ $\mathcal{L}^m_{cls}$ and $\mathcal{L}^m_{fusion}$, respectively. Besides, the deviation between predictions and ground truth is optimized by $\mathcal{L}_{gt}$. During training, the unimodal momentum models are updated in a momentum manner and utilized to provide pseudo targets for momentum distillation.

Furthermore, by assigning learnable weight factors of the modalities in the stage of fusion, *Multimodal Fusion Distillation (MFD)* adjusts the contribution of three modalities and promotes efficient interaction of the modality-share features.

However, despite the effectual knowledge distillation from two teacher networks to student networks, the textual modality has subjective and biased emotion issues as declared in the introduction while the multimodal subtask suffers from the unstable inference problem due to the randomly initialized interaction layers. These troubles cause the partial unreliability of teacher networks, which may provide wrong ground truth to the unimodal subtask. Moreover, there is a risk of unimodal subtasks over-reliance on the teacher networks, resulting in the neglect of the modality-specific features. Inspired but diverse from ALBEF [59], to reduce the penalties on the predictions inferred by unimodal subtask regardless of the correctness and modality-specific information, we introduce *momentum distillation* for the unimodal and multimodal subtasks.

- In the training stage, for acoustic and visual modalities, we construct *unimodal momentum model* as a continuously-evolving teacher network that consists of exponential-moving-average counterparts of audio and vision encoders with their prediction heads. Generated by the unimodal momentum model, the pseudo-targets including unimodal features and prediction labels, are employed to train the unimodal subtasks jointly with the base-outputs from dominant modality and multimodal distillation. The momentum models encourage the unimodal subtasks to focus on modality-specific features, whose effect on the final sentiment prediction can be adaptively balanced with modality-shared features in the proposed MTMD.
- For the multimodal fusion subtask, the learnable weight factors of different modalities in the multimodal representation are updated similarly in the exponential-moving-average way, named as *momentum fusion factors*. The momentum learning strategy of the weight factors makes the contribution adjustment of different modalities more steadily, leading to a more robust multimodal representation.

C. Unimodal Subtasks Setup

We first introduce the network setup of unimodal subtasks for the utterance inputs $X_u \in \mathbb{R}^{\ell_u \times d_u}$, where $u \in \{t, a, v\}$. For textual modality, we use pre-trained BERT [10] to encode the text inputs X_t with attention mask and segment IDs into textual representation F'_t, where the output embeddings from the last Transformer layer can be represented as:

$$F'_t = BERT(X_t; \theta_t) \in \mathbb{R}^{\ell_t \times d_t} \tag{1}$$

For acoustic and visual modalities, we use Toolkits as the unimodal extractor to extract the original unimodal features, denoted as X_a and X_v, which are the inputs to the acoustic and visual encoder. As shown in Fig. 2, the unimodal encoders consist of a single directional Long Short-Term Memory (sLSTM) [62]

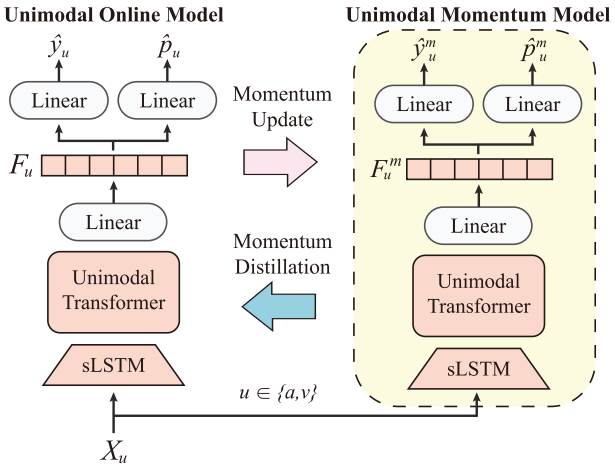


Fig. 2. Illustration of the unimodal online model and unimodal momentum model for acoustic and visual modalities. As the exponential-moving-average counterparts of the corresponding online models, the momentum models provide momentum-updated representations and predictions as the pseudo-targets in momentum distillation.

to capture the temporal characteristics and a designed 3-layer Transformer [63] to encode the global attention information. For $u \in \{a, v\}$, the outputs of unimodal encoders are represented as:

$$h_u = sLSTM(X_u; \theta_u) \in \mathbb{R}^{\ell_u \times d_u}$$
$$F'_u = Transformer(h_u; \theta_u) \in \mathbb{R}^{\ell_u \times d_u} \quad (2)$$

Note that we only utilize the [CLS] token of F'_t and the embedding from the first time step of F'_a and F'_v, meaning that F'_u in the following process satisfies $F'_u \in \mathbb{R}^{d_u}$ for $u \in \{t, a, v\}$. Therefore, the proposed approach is suitable for modality inputs with various length sequences and capable of dealing with both token-aligned and -unaligned multimodal data.

In order to better interact and distill knowledge among the unimodal subtasks, we project the unimodal representation into a new feature space with the same feature dimension. After the linear projection with a $ReLU$ activation layer, we obtain the final representations of three modalities which are employed for further distillation of unimodal subtasks, represented as:

$$F_u = ReLU(W_u F'_u + b_u) \in \mathbb{R}^d, \ u \in \{t, a, v\} \quad (3)$$

For the sentiment prediction of three unimodal subtasks, we apply a regression head and a classification head to predict the sentiment score and intensity, respectively:

$$Regression: \hat{y}_u = ReLU(W^r_u F_u + b^r_u) \in \mathbb{R}^1$$
$$Classification: \hat{p}_u = ReLU(W^c_u F_u + b^c_u) \in \mathbb{R}^3 \quad (4)$$

where $u \in \{t, a, v\}$, the regression head outputs a continuous value denoting the sentiment score of the corresponding modality, and the classification head outputs three values denoting the probability of three sentiment class in $\{positive, neutral, negative\}$.

With the ground truth labels denoted as y and p, we formulate $\mathcal{L}^u_{gt}$ summed by Mean Absolute Error (MAE) loss and Cross-Entropy loss as:

$$\mathcal{L}^u_{gt} = MAE(y, \hat{y}_u) + CrossEntropy(p, \hat{p}_u)$$
$$= \frac{1}{N}\sum_{i=1}^{N} |y^i - \hat{y}^i_u| + \frac{1}{N}\sum_{i=1}^{N} H(p^i, \hat{p}^i_u) \quad (5)$$

where N denotes the number of training samples.

D. Unimodal Momentum Model

Concurrently, we retain the momentum versions of the audio and vision encoders with the prediction heads, named unimodal momentum models for both acoustic and visual modalities. Similar to ALBEF [59], the unimodal momentum models are designed as the exponential-moving-average counterparts of the corresponding online models, whose parameters can be represented as:

$$para_i = \eta \cdot para_i + (1 - \eta) \cdot para_{i-1} \quad (6)$$

where $para$ denotes the parameters of the unimodal momentum model and η denotes the momentum-updated factor, which is set as 0.995 in our paper.

As shown in Fig. 2, for $u \in \{a, v\}$, unimodal momentum models generate unimodal representations $F^m_u \in \mathbb{R}^d$, regression prediction labels $\hat{y}^m_u \in \mathbb{R}^1$ and classification prediction labels $\hat{p}^m_u \in \mathbb{R}^3$, which constitute the pseudo-targets utilized in momentum distillation.

E. Dominant Modality Distillation (DMD)

To utilize semantic information extracted in the textual subtask, we regard textual modality as the dominant modality and present dominant modality distillation on acoustic and visual subtasks.

As CRD [56] shows that combining the original knowledge distillation [51] and the representation knowledge distillation greatly improve the transferability of the learned model, we construct both response-based and feature-based DMD across the unimodal subtasks. The former utilizes response-based knowledge distillation while the latter constructs feature-based contrastive learning for distillation, as shown in Fig. 3.

1) Response-based Knowledge Distillation aims at deeming the prediction of textual subtask as base-outputs which are considered to be the soft ground truth label in DMD. In the training of the proposed MTMD, we observe that the outputs of textual subtask can reach an accuracy close to 100%, which is deduced to be caused by the overfitting problem. Instead of directly utilizing the overfitting outputs of textual subtask as the final prediction, we deem the textual outputs as teacher predictions whose knowledge is worthy for the subtasks of inferior modalities to attain more informative supervised signals. Moreover, with the additional supervision of base-outputs as soft labels, the proposed approach further improve the optimization efficiency of unimodal subtasks.

Since we take textual modality as the dominant modality, the outputs of textual subtask $\hat{y}_t$ and $\hat{p}_t$ are seen as teacher labels

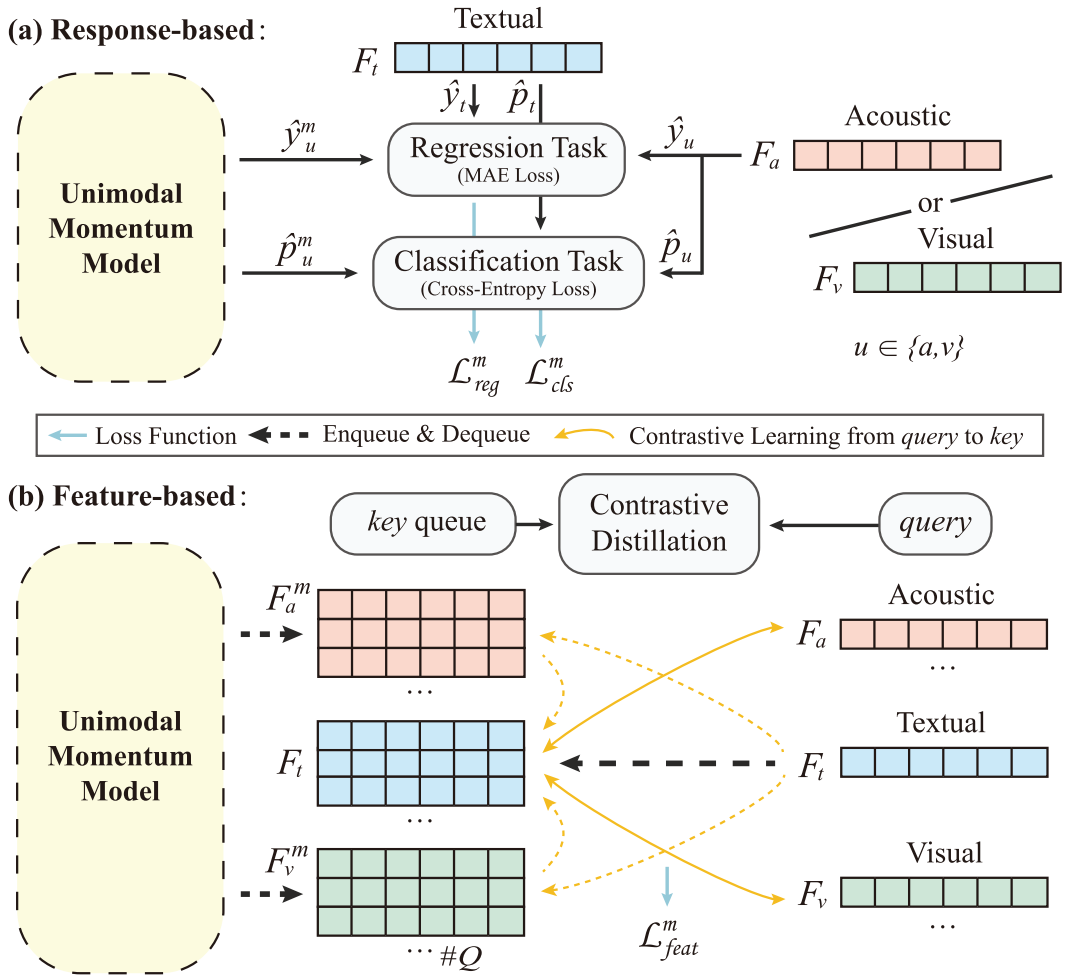


Fig. 3. Illustration of the Dominant Modality Distillation (DMD), divided into (a) response-based and (b) feature-based knowledge distillation, optimized by $\mathcal{L}^m_{reg}/\mathcal{L}^m_{cls}$ and $\mathcal{L}^m_{feat}$ respectively. Deeming textual modality as the dominant modality, the former transfers the regression predictions $\hat{y}_t$ and classification predictions $\hat{p}_t$ of textual modality to the inferior modalities $\hat{y}_u/\hat{p}_u$ for $u \in \{a, v\}$; the latter constructs contrastive learning between the representations of textual modality F_t and the other two modalities F_a/F_v with the maintained queue containing the most recent unimodal representations. For acoustic and visual modalities, the unimodal momentum model generates the corresponding predictions $\hat{y}^m_u/\hat{p}^m_u$ and representations F^m_u as pseudo-targets to guide the transfer of knowledge and explore the modality-specific features for inferior modalities.

to acoustic and visual subtasks. According to the regression and classification tasks, we use MAE loss and Cross-Entropy loss to formulate the response-based knowledge distillation loss $\mathcal{L}^u_{reg}$ and $\mathcal{L}^u_{cls}$ for $u \in \{a, v\}$, represented as:

$$\mathcal{L}^u_{reg} = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}^i_t - \hat{y}^i_u|$$

$$\mathcal{L}^u_{cls} = \frac{1}{N}\sum_{i=1}^{N} H(\hat{p}^i_t, \hat{p}^i_u) \tag{7}$$

2) Feature-based Knowledge Distillation concentrates on extracting semantic and sentiment-related features of acoustic and visual modalities in the guidance of textual representations with the same sentiment class, which is divided into $\{positive, neutral, negative\}$. Contrastive learning is conducted in feature-based knowledge distillation to deeply interacts the information among different modality representations and capture the inter-modal dynamics. Diverse from ALBEF [59] trained on unsupervised data, we introduce the sentiment classes to supervise the contrastive learning among different modalities, which promotes the feature-based knowledge distillation to learn discriminative representation for different polarities of sentiments.

Specifically, given the representations $\{F_t, F_a, F_v\}$ from different modalities, we view textual representation F_t as teacher representation while the representation of inferior modalities F_a and F_v as student representations. In the cross-modal contrastive learning [60], the interaction among encoded representations is deemed as a matching task for query and key samples from different modalities respectively, whose goal is to increase the similarity between matched positive samples while decrease the similarity between query and other negative key samples. Inspired by MoCo [64], we maintain queues of key samples for each modality to store the most recent Q representations output from the unimodal online model and momentum model. In order to pull closer the representations from the same sentiment class while push apart the ones from different classes,

we calculate cross-modal similarity functions $s(t,u) = \overline{F}_t^{\mathrm{T}} \overline{F}_u$ and $s(u,t) = \overline{F}_u^{\mathrm{T}} \overline{F}_t$, where $\overline{F}_t$ and $\overline{F}_u$ are the L2-normalized representations of corresponding modalities.

Then for textual and the other two inferior modalities of ith sample, following SupCon [65], the softmax-normalized similarity is computed as:

$$V_i^{t2u}(t) = \frac{\sum_{k=1}^{K} \exp(s(t,u_k)/\tau)}{\sum_{i=1}^{N} \exp(s(t,u_i)/\tau)}$$
$$V_i^{u2t}(u) = \frac{\sum_{k=1}^{K} \exp(s(u,t_k)/\tau)}{\sum_{i=1}^{N} \exp(s(u,t_i)/\tau)} \tag{8}$$

where τ is a learnable temperature parameter regulating the probability distribution over distinct instances [51] and K denotes the number of positive samples in the same sentiment class. Note that the keys u_k/u_i in $V_i^{t2u}(t)$ and t_k/t_i in $V_i^{u2t}(u)$ are online unimodal representations from the maintained queues.

Let $Y^{t2u}(t)$ and $Y^{u2t}(u)$ denote the ground-truth one-hot similarity, where representations come from the same sentiment class have a probability of 1 and representations come from different sentiment classes have a probability of 0. For $u \in \{a,v\}$, the feature-based contrastive distillation loss $\mathcal{L}_{feat}^{u}$ is formulated as the Cross-Entropy loss between Y and V:

$$\mathcal{L}_{feat}^{u} = \frac{1}{2} \mathop{\mathbb{E}}_{t,u} \left[H(Y^{t2u}(t), V^{t2u}(t)) + H(Y^{u2t}(u), V^{u2t}(u)) \right]$$
$$= \frac{1}{2N} \left[\sum_{i=1}^{N} H(Y_i^{t2u}(t), V_i^{t2u}(t)) + \sum_{i=1}^{N} H(Y_i^{u2t}(u), V_i^{u2t}(u)) \right] \tag{9}$$

Momentum Distillation: To improve knowledge distillation under implicit noisy supervision of dominant modality during training, we propose to learn from pseudo-targets generated by the unimodal momentum model, including pseudo labels $\hat{y}_u^m$, $\hat{p}_u^m$ and pseudo features F_u^m for $u \in \{a,v\}$. Specifically, the momentum distillation loss is a weighted combination of the original task's loss and the MAE or Kullback-Leibler (KL) divergence loss between the pseudo-targets and the model's prediction.

By matching the predictions from the unimodal online model with the unimodal momentum model, the final response-based momentum distillation loss $\mathcal{L}_{reg}^{m}$ and $\mathcal{L}_{cls}^{m}$ is defined as:

$$\mathcal{L}_{reg}^{m} = \sum_{u \in \{a,v\}} \left[(1-\alpha)\mathcal{L}_{reg}^{u} + \alpha \frac{1}{N} \sum_{i=1}^{N} |(\hat{y}_u^m)^i - \hat{y}_u^i| \right]$$
$$\mathcal{L}_{cls}^{m} = \sum_{u \in \{a,v\}} \left[(1-\alpha)\mathcal{L}_{cls}^{u} + \alpha \frac{1}{N} \sum_{i=1}^{N} H((\hat{p}_u^m)^i, \hat{p}_u^i) \right] \tag{10}$$

where α denotes the weighted contribution of the unimodal momentum model.

For feature-based momentum distillation, we first utilize the representations F_u^m output by unimodal momentum model to compute similarity functions as $s^m(t,u^m) = \overline{F}_t^{\mathrm{T}} \overline{F_u^m}$ and $s^m(u^m,t) = \overline{F_u^m}^{\mathrm{T}} \overline{F}_t$. By replacing s with s^m in (8), we compute the soft pseudo-targets $U^{t2u}(t)$ and $U^{u2t}(u^m)$. Note that in momentum distillation, the keys for $u \in \{a,v\}$ are the maintained queues generated from the unimodal momentum model, while the keys for t are from queues containing online textual representations. Accordingly, the final feature-based momentum distillation loss $\mathcal{L}_{feat}^{m}$ is defined as:

$$\mathcal{L}_{feat}^{m} = \sum_{u \in \{a,v\}} \{(1-\alpha)\mathcal{L}_{feat}^{u} + \frac{\alpha}{2} \mathop{\mathbb{E}}_{t,u} \left[KL(U^{t2u}(t) \parallel V^{t2u}(t)) + KL(U^{u2t}(u^m) \parallel V^{u2t}(u)) \right] \} \tag{11}$$

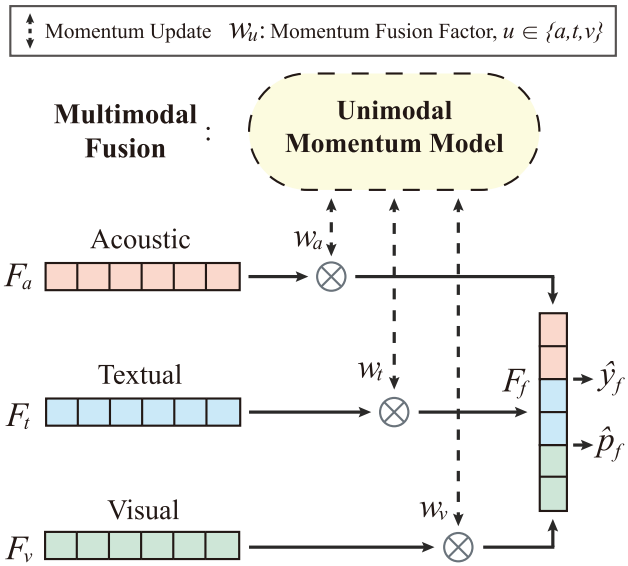


Fig. 4. Illustration of the Multimodal Fusion. In the concatenation to acquire the final multimodal representations, the unimodal representations F_u from different modalities $u \in \{t,a,v\}$ are respectively weighted by the fusion factors $\{w_t, w_a, w_v\}$, which are momentum-updated as the unimodal momentum model. After projecting, the multimodal representations F_f are fed into regression and classification head to infer the final predictions $\hat{y}_f$ and $\hat{p}_f$ in the multimodal subtask.

F. Multimodal Fusion Distillation (MFD)

To obtain a robust multimodal representation, we adjust the contribution of different modalities according to the effectiveness of the corresponding unimodal representations in a learnable way during multimodal fusion, as shown in Fig. 4. Specifically, after obtaining the final unimodal representations F_u as shown in (3), we multiply them with momentum fusion factors denoted as $\{w_t, w_a, w_v\}$ which adaptively adjust the contribution of different modalities in the multimodal representation. Then we concatenate each representation and utilize a MLP to interact the features of unimodal representation. The final multimodal representation F_f is formulated as:

$$F_f' = Concat(w_t F_t; w_a F_a; w_v F_v)$$
$$F_f = ReLU(W^f F_f' + b^f) \in \mathbb{R}^{d_f} \tag{12}$$

where d_f is the feature dimension of the multimodal representation. To avoid the instability brought by the randomly initialized MLP, the learnable weight factors are updated in a momentum-updated way as the parameters of the unimodal momentum model do in (6), leading to a more stable fusion stage.

Similar to the unimodal subtasks, we utilize a regression head and a classification head to the multimodal representation F_f in the multimodal subtask and obtain the predicted regression labels $\hat{y}_f$ and classification labels $\hat{p}_f$. In addition, with the ground truth labels p and y, the training loss of the multimodal subtask is formulated as:

$$\begin{aligned}\mathcal{L}_{gt}^{f} &= MAE(y, \hat{y}_f) + CrossEntropy(p, \hat{p}_f) \\ &= \frac{1}{N}\sum_{i=1}^{N}|y^i - \hat{y}_f^i| + \frac{1}{N}\sum_{i=1}^{N}H(p^i, \hat{p}_f^i)\end{aligned} \tag{13}$$

Momentum Distillation: Intuitively, the multimodal representation contains the interaction among different modalities, being more informative than the individual unimodal representations. Thus, the predictions $\hat{y}_f$ and $\hat{p}_f$ output from the multimodal subtask can also be observed as the teacher network to the acoustic and visual subtasks as textual subtask does. The knowledge distillation among multimodal subtask and unimodal subtasks can further encourage the unimodal online models to capture the inter-modal dynamics which are beneficial for multimodal fusion. Moreover, due to the abundant semantic information of multimodal subtask which constructs more close relationship with the ground truth sentiment, we propose a hard samples attention strategy in the loss function to promote the knowledge distillation among multimodal and unimodal subtasks. Concretely, the hard samples attention strategy assigns more weights for the samples with more variation between multimodal predictions and the ground truth labels and assigns fewer weights for the samples with more accurate predictions. For regression and classification subtasks, the corresponding loss weights of samples are computed as:

$$\begin{aligned}Regression &: w_{reg}^i = \tanh|y^i - \hat{y}_f^i| \\ Classification &: w_{cls} = \frac{1}{N}\sum_{i=1}^{N}|p^i - argmax(\hat{p}_f^i)|\end{aligned} \tag{14}$$

With multimodal predictions as teacher supervisions while unimodal predictions as student learners, the weighted multimodal regression and classification knowledge distillation loss is represented as:

$$\begin{aligned}\mathcal{L}_{fusion}^{u} &= w_{reg}MAE(\hat{y}_f, \hat{y}_u) + w_{cls}CrossEntropy(\hat{p}_f, \hat{p}_u) \\ &= \frac{1}{N}\sum_{i=1}^{N}w_{reg}^i \cdot |\hat{y}_f^i - \hat{y}_u^i| + w_{cls}\cdot\frac{1}{N}\sum_{i=1}^{N}H(\hat{p}_f^i, \hat{p}_u^i)\end{aligned} \tag{15}$$

The final momentum fusion distillation loss $\mathcal{L}_{fusion}^{m}$ is defined as:

$$\mathcal{L}_{fusion}^{m} = \sum_{u\in\{a,v\}} L_{fusion}^{u} \tag{16}$$

Algorithm 1: Multi-Task Momentum Distillation.

Input: unimodal features X_u, $u \in \{t, a, v\}$ with the corresponding ground truth labels y and p; hyper-parameters α, τ, $\gamma_1 \sim \gamma_4$
Output: multimodal sentiment prediction $\hat{y}_f$

Initialize unimodal online and momentum models for DMD and multimodal fusion model for MFD;
Create queues of unimodal representations in buffer;
for *each utterance sample in the mini-batch* **do**
  (1) Unimodal subtasks;
  for *each modality* $u \in \{t, a, v\}$ **do**
    Obtain final unimodal representations F_u;
    Obtain unimodal prediction $\hat{y}_u$ and $\hat{p}_u$;
    Compute $\mathcal{L}_{gt}^{u}$ as in Eq.5;
    Compute $\mathcal{L}_{reg}^{u}$ and $\mathcal{L}_{cls}^{u}$ as in Eq.7;
    Compute $\mathcal{L}_{feat}^{u}$ as in Eq.9;
  end
  Update parameters of unimodal momentum models as in Eq.6;
  for *each modality* $u \in \{a, v\}$ **do**
    Obtain pseudo-targets including features F_u^m and labels $\hat{y}_u^m$, $\hat{p}_u^m$;
    Conduct momentum distillation in DMD and compute $\mathcal{L}_{reg}^{m}$, $\mathcal{L}_{cls}^{m}$, $\mathcal{L}_{feat}^{m}$ as Eq.10-11;
  end
  (2) Multimodal subtask;
  Calculate momentum fusion factors w_t, w_a, w_v;
  Obtain final multimodal representations F_f;
  Obtain multimodal prediction $\hat{y}_f$ and $\hat{p}_f$;
  Compute $\mathcal{L}_{gt}^{f}$ as in Eq.13;
  Conduct momentum distillation in MFD and compute $\mathcal{L}_{fusion}^{m}$ as Eq.15-16;
  (3) Total Optimization;
  Compute total training loss $\mathcal{L}_{total}$ as Eq.18;
  Compute parameters gradient $\partial\mathcal{L}_{total}/\partial\theta$;
  Update model parameters except for unimodal momentum models;
end

G. Optimization Objective

Combing unimodal and multimodal subtasks, the ground truth loss $\mathcal{L}_{gt}$ is represented as:

$$\mathcal{L}_{gt} = \sum_{u\in\{t,a,v\}} \mathcal{L}_{gt}^{u} + \mathcal{L}_{gt}^{f} \tag{17}$$

Considering $\mathcal{L}_{reg}^{m}$ and $\mathcal{L}_{cls}^{m}$ from the response-based dominant modality momentum distillation, $\mathcal{L}_{feat}^{m}$ from the feature-based dominant modality momentum distillation and L_{fusion}^{m} from the multimodal fusion momentum distillation, the total training loss is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{gt} + \gamma_1\mathcal{L}_{reg}^{m} + \gamma_2\mathcal{L}_{cls}^{m} + \gamma_3\mathcal{L}_{feat}^{m} + \gamma_4\mathcal{L}_{fusion}^{m} \tag{18}$$

where $\gamma_1 \sim \gamma_4$ are weighted hyper-parameters to adjust the impact of various losses from different subtasks.

The training pipeline of the proposed MTMD is illustrated in Algorithm 1, which summarizes the whole architecture including DMD and MFD as pseudo-code.

IV. Mutual Information Maximization Perspective

In this section, we offer an alternative perspective of the proposed MTMD to show that the approach maximizes a lower bound on the mutual information (MI) among different views of multiple subtasks and modalities. The presented ideas of multi-task learning and knowledge momentum distillation can be interpreted as generating sentiment-related views in different ways.

Formally, we define two different views as T and S representing two subtasks or two modalities in the proposed approach. We consider learning representations sentiment-invariant and modality-invariant to the change of views, which can be achieved by maximizing the MI between T and S.

Response-Based Knowledge Distillation: We denote T as the teacher network while S as the student network, and maximize a lower bound on $MI(T,S)$ formulated as Agakov and Barber [66]:

$$\begin{aligned} MI(T,S) &= H(T) - H(T \mid S) \\ &= H(T) + \mathop{\mathbb{E}}_{T,S}[\log P(T \mid S)] \\ &= H(T) + \mathop{\mathbb{E}}_{T,S}[\log Q(T \mid S)] + \mathop{\mathbb{E}}_{S}[KL(P(T \mid S) \parallel Q(T \mid S))] \\ &\geq H(T) + \mathop{\mathbb{E}}_{T,S}[\log Q(T \mid S)] \end{aligned} \tag{19}$$

where $MI(T,S)$ can be considered as the shared knowledge of the teacher network and the student network while $P(T \mid S)$ and $Q(T \mid S)$ represent the true distribution and approximate distribution of the transferable knowledge. The last inequality holds due to the non-negativity of the KL divergence. $\mathcal{L}^u_{reg}$ and $\mathcal{L}^u_{cls}$ aim at minimizing the distribution distance between the prediction of dominant modality and inferior modality, i.e., maximizing $\mathbb{E}_{T,S}[\log Q(T \mid S)]$. Similarly, $\mathcal{L}^m_{fusion}$ focuses on matching the prediction distribution between multimodal and unimodal subtasks. These response-based losses maximize the MI among different modalities on the level of sentiment prediction and interact the predicted label distribution of different subtasks.

Feature-Based Knowledge Distillation: We represent T as the teacher modality while S as the student modality. Diverse from ALBEF [59] which analyzes unsupervised contrastive learning losses, we introduce sentiment-class labels as supervisions to the contrastive learning of teacher and student modalities in feature-based knowledge distillation to further maximize sentiment-related information for downstream sentiment analysis. Following CPC [67] and SupCon [65], we maximize a lower bound on $MI(T,S)$ by minimizing the InfoNCE loss, which is defined as:

$$MI(T,S) \geq \log(N) - \mathcal{L}_{NCE} \tag{20}$$

$$\mathcal{L}_{NCE} = -\frac{1}{|D_k(S)|} \mathop{\mathbb{E}}_{T,S}\left[\log \frac{\sum_{S_k \in D_k(S)} \exp(s(T,S_k))}{\sum_{S_i \in D(S)} \exp(s(T,S_i))}\right] \tag{21}$$

where $D(S)$ contains $|D_k(S)|$ positive sample S in the same sentiment classes and $|D(S) \backslash D_k(S)|$ negative samples drawn from the proposal distribution of samples in different sentiment classes, and $s(T,S)$ is a scoring function to measure the similarity between T and S, which denotes the dot-product between two representations in our paper.

The feature-based contrastive distillation loss L^u_{feat} with one-hot labels $Y(t,u)$ in (9) can be rewritten as:

$$\begin{aligned} L^u_{feat} = -\frac{1}{2K} \times \mathop{\mathbb{E}}_{Y}\Bigg[&\log \frac{\sum_{k=1}^{K} \exp(s(t,u_k)/\tau)}{\sum_{i=1}^{N} \exp(s(t,u_i)/\tau)} \\ &+ \log \frac{\sum_{k=1}^{K} \exp(s(u,t_k)/\tau)}{\sum_{i=1}^{N} \exp(s(u,t_i)/\tau)}\Bigg] \end{aligned} \tag{22}$$

where minimizing L^u_{feat} is equivalent to minimizing the symmetric version of InfoNCE, which further achieves the goal to maximize the MI between the dominant modality t and the inferior modality $u \in \{a,v\}$. The feature-based contrastive distillation loss further promotes the model to extract the modality-shared features and learn the modality-invariant representations among different modalities.

MI on Multiple Views: Both response-based and feature-based knowledge distillation can be considered as generating multiple views by either the dominant modality subtask or multimodal subtask to the prediction layer or feature layer of the unimodal subtasks. The momentum distillation and the setting of regression and classification heads can be seen as generating different views for the sentiment task.

1) Momentum distillation generates alternative views from the proposal distribution for predictions or representations of the unimodal subtasks. Taking L^m_{feat} in (11) as an example, minimizing $KL(U^{u2t}(u^m) \parallel V^{u2t}(u))$ is equivalent as minimizing the following objective:

$$\begin{aligned} &-\sum_k U^{u2t}_k(u^m) \log V^{u2t}_k(u) \\ &= -\sum_k \frac{\exp(s^m(u^m,t_k)/\tau)}{\sum_{i=1}^{N} \exp(s^m(u^m,t_i)/\tau)} \log \frac{\exp(s(u,t_k)/\tau)}{\sum_{i=1}^{N} \exp(s(u,t_i)/\tau)} \end{aligned} \tag{23}$$

which promotes $s(u,t_k)$ to match $s^m(u^m,t_k)$ and tends to preserve the modality-specific information when maximizing $MI(u,t_k)$.

2) The regression and classification subtasks can be considered as different views of the sentiment prediction task, which utilize the same representations to output sentiment scores and emotion classes, respectively. By maximizing the MI on different sentiment-related subtasks for each sample, the learned representations can capture the sentiment-invariant semantic knowledge among the subtasks. Furthermore, the setting of different task-specific heads belongs to the hard parameter sharing method in multi-task learning [41], which lowers the risk of overfitting to one subtask, and encourages the unimodal and multimodal subtasks to learn more robust and universal representations.

V. Experiments and Discussion

A. Datasets and Evaluation Metric

CMU-MOSI [68] is a public-used benchmark dataset in the research field of MSA, which contains 2,199 monologue utterances segmented from 93 opinion videos taken by 89 YouTube movie reviewers. Each utterance is manually annotated with a continuous sentiment score from −3 (strongly negative) to +3 (strongly positive). We utilize 1,284 utterances for training, 229 utterances for validation, and 686 utterances for testing.

CMU-MOSEI [69] expands the size of the multimodal dataset into 3,228 videos in 250 diverse topics collected by 1,000 distinct YouTube speakers. The dataset consists of 20 k video clips, each of which is annotated for the sentiment on a [−3, +3] Likert scale and Ekman emotion [70] classes of $\{happiness, sadness, anger, fear, disgust, surprise\}$. We utilize 16,326 utterances for training, 1,871 utterances for validation, and 4,659 utterances for testing.

The evaluation metrics are set the same for both of the datasets in the MSA task. For sentiment intensity classification, we utilize seven-class classification accuracy (Acc7) to measure the correct sentiment prediction classes ranging from [−3, +3]. Moreover, we report binary classification accuracy (Acc2) and Weighted F1-score (F1) in two calculation settings as negative &non-negative (non-exclude 0) [1] / negative& positive (exclude 0) [2]. For sentiment score regression, we calculate mean absolute error (MAE) for the difference between ground truth and prediction labels, and Pearson correlation (Corr) for the degree of prediction skew.

B. Implementation Details

We introduce the pre-processing methods especially the unimodal feature extractors, and the experimental details including the hardware environment and selection of hyper-parameters in the following.23

- *Textual:* As said in Section III-C, we utilize the pre-trained BERT-base-uncased model to encode the text input with word attention mask and sentence segment IDs. The textual features have 768-dimensional hidden states for each word token.
- *Acoustic & Visual:* As mentioned by Mao et al. [12], due to the vague description of extractor backbone and feature selection, we directly adopt the CMU-Multimodal SDK [31] to pre-process the audio and vision data following [11], [30], [71], [72]. The SDK applies COVAREP [73] to extract the acoustic features including 12 Mel-frequency cepstral coefficients (MFCCs), pitch tracking, speech polarity, spectral envelope, glottal source parameters, peak slope parameters, and maxima dispersion quotients, etc., and implements Facet [74] to extract the visual features including facial landmarks, head pose, action units, gaze tracking and histogram of oriented gradients (HOG) features, etc.

We conduct the experiments on a single GTX 1080Ti GPU with CUDA 10.2. For the hyper-parameters, following Gkoumas et al. [75], we perform fifty-times random grid search to fine-tune the model including $\gamma_1 \sim \gamma_4$ in $\{0.6, 0.8, 1.0\}$, α in $\{0.2, 0.3, 0.4, 0.5, 0.6\}$ and τ in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. The batch sizes for CMU-MOSI and CMU-MOSEI are set as 64 and 128, respectively. According to the scale of datasets, we set the size of representation queues in feature-based contrastive distillation as 1,024 on CMU-MOSI and 8,192 on CMU-MOSEI. The dropout value is set as 0.1. For optimization, we adopt Adam [76] as the optimizer with the learning rate 5e-5 for the parameters of BERT, 5e-3 for the parameters of the unimodal online model, and 1e-3 for the other parameters. The sLSTM is 1-layer with 32 and 64 as the output dimension for acoustic and visual modalities, correspondingly. The acoustic and visual Transformers have 3 stacked layers with 4 parallel attention heads in each layer. We run experiments with the best hyper-parameters setting five times and report the average performance as the final results.

C. Baselines

To fairly compare the performance between the MSA baselines and our proposed approach, we utilize the pre-trained BERT language model as the textual features encoder and the CMU-Multimodal SDK as the acoustic and visual features extractor to reproduce the baselines, remaining the same as our approach. The results of baselines are reproduced using publicly available source codes and the same setting of hyper-parameters as the original papers. The baselines are introduced in detail as below:

EF-LSTM (Early Fusion LSTM) [29] concatenates the modality features at the input level and processes the concatenated features with a LSTM layer and classifier.

LF-DNN (Late Fusion Deep Neural Network) [36] utilizes three separate deep neural networks to encode features for corresponding modalities and then combine them to make the final decision.

TFN (Tenser Fusion Network) [1] explicitly computes unimodal, bimodal, and multimodal tensors to capture different levels of inter-modal dynamics across three modalities.

LMF (Low-rank Multimodal Fusion) [30] reduces the complexity of multimodal tensors by low-rank factors to decompose the stacked high-order tensors.

MFN (Memory Fusion Network) [35] individually models view-specific and cross-view interactions temporally by delta-memory attention network and utilizes a multi-view gated memory to summarize.

Graph-MFN (Graph Memory Fusion Network) [69] replaces the fusion components in MFN with a dynamic fusion graph module to learn the fusion mechanism of different combinations of three modalities.

MFM (Multimodal Factorization Model) [13] disentangles the intra-modal and cross-modal interactions into multimodal discriminative factors and modality-specific generative factors in multimodal learning.

MulT (Multimodal Transformer) [2] extends the Transformer architecture by directional pairwise cross-modal attention to learn the multimodal representation.

MISA (Modality-Invariant and -Specific Representations) [3] projects features into modality-invariant and modality-specific

TABLE I
PERFORMANCE COMPARISON BETWEEN MTMD AND BASELINES ON CMU-MOSI AND CMU-MOSEI DATASETS

Models	CMU-MOSI					CMU-MOSEI				
	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑
EF-LSTM [29]	34.5	77.8 / 79.0	77.7 / 78.9	0.952	0.651	49.3	80.1 / 80.3	80.3 / 81.0	0.603	0.682
LF-DNN [36]	33.6	78.0 / 79.3	77.9 / 79.3	0.978	0.658	52.1	78.6 / 82.3	79.0 / 82.2	0.561	0.723
TFN [1]	33.7	78.3 / 80.2	78.2 / 80.1	0.925	0.662	52.2	81.0 / 82.6	81.1 / 82.3	0.570	0.716
LMF [30]	32.7	77.5 / 80.1	77.3 / 80.0	0.931	0.670	52.0	81.3 / 83.7	81.6 / 83.8	0.568	0.727
MFN [35]	34.2	77.9 / 80.0	77.8 / 80.0	0.951	0.665	51.1	81.8 / 84.0	81.9 / 83.9	0.575	0.720
Graph-MFN [69]	34.4	77.9 / 80.2	77.8 / 80.1	0.939	0.656	51.9	81.9 / 84.0	82.1 / 83.8	0.569	0.725
MFM [13]	33.3	77.7 / 80.0	77.7 / 80.1	0.948	0.664	50.8	80.3 / 83.4	80.7 / 83.4	0.580	0.722
MulT [2]	35.0	79.0 / 80.5	79.0 / 80.5	0.918	0.685	52.1	81.3 / 84.0	81.6 / 83.9	0.564	0.732
MISA [3]	43.5	81.8 / 83.5	81.7 / 83.5	0.752	0.784	52.2	81.6 / 84.3	82.0 / 84.3	0.550	0.758
MAG-BERT [11]	45.1	82.4 / 84.6	82.2 / 84.6	0.730	0.789	52.8	81.9 / 85.1	82.3 / 85.1	0.558	0.761
Self-MM [19]	45.8	82.7 / 84.9	82.6 / 84.8	0.731	0.785	53.0	82.6 / 85.2	82.8 / 85.2	0.540	0.763
MMIM [16]	45.0	83.0 / 85.1	82.9 / 85.0	0.738	0.781	53.1	81.9 / 85.1	82.3 / 85.0	0.547	0.752
MTMD	**47.5**	**84.0 / 86.0**	**83.9 / 86.0**	**0.705**	**0.799**	**53.7**	**84.8 / 86.1**	**84.9 / 85.9**	**0.531**	**0.767**

subspaces to distinctly learn the commonalities and characteristic features of different modalities.

MAG-BERT (Multimodal Adaption Gate for BERT) [11] dynamically shifts the verbal feature by nonverbal information to fine-tune the internal representations of the vanilla BERT model.

Self-MM (Self-Supervised Multi-Task Learning) [19] assigns the automatically generated labels to each modality for better adjustment in the gradient back-propagation.

MMIM (Multimodal Mutual Information Maximization) [16] maximizes the mutual information among the unimodal input pairs and the multimodal fusion output to extract the task-related information.

D. Experiment Results

In this section, we compare the proposed MTMD with other baselines on CMU-MOSI [68] and CMU-MOSEI [69], and conduct the ablation study to demonstrate the effectiveness of the proposed method.

1) Evaluation on CMU-MOSI Dataset: As shown in Table I, the proposed MTMD surpasses the MSA baselines on all metrics on CMU-MOSI datasets. Specifically, for the previous methods, Self-MM [19] beats other baselines on Acc7 (45.8%) and MAG-BERT [11] performs better on MAE (0.730) and Corr (0.789), while MMIM [16] attains the state-of-the-art performance on Acc2 (83.0%/85.1%) and F1 (82.9%/85.0%). We suppose that the unimodal labels generated by Self-MM are more precisely to express the minor differences among multiple intensity classes of sentiment while the non-verbal shifting by MAG-BERT on the verbal representations can be seen as a fine-tuning to compute more accurately continuous values of regression. In addition, MMIM focuses on the mutual information among different modality representations which is more likely to capture the commonalities of sentiment from the same positive or negative classes. Diverse from the previous methods, the proposed MTMD distills knowledge at both label and feature levels by the response-based and feature-based momentum distillation, which gives consideration to both inter-modal and intra-modal dynamics. Therefore, MTMD outperforms Self-MM by 1.7% on Acc7 and exceeds MAG-BERT by 0.025 on MAE and 0.01 on Corr. Moreover, MTMD surpasses MMIM by 1.0%/0.9% on Acc2 and 1.0%/1.0% on F1.

2) Evaluation on CMU-MOSEI Dataset: Similarly, the proposed MTMD achieves better performance than state-of-the-art methods on CMU-MOSEI dataset as shown in Table I. In a larger scale of dataset, MMIM catches up with and even slightly outmatches Self-MM on Acc7 (53.1%). However, Self-MM obtains the state-of-the-art performance on other metrics, including Acc2 (82.6%/85.2%), F1 (82.8%/85.2%), MAE (0.540) and Corr (0.763), which is observed that based on more data, the unimodal labels can be generated more stably. Nevertheless, the hard label supervision limits the information abundance and is easily interfered by data outlier noises. On the contrary, we utilize soft label supervision as the pseudo-targets generated by the momentum unimodal model to encourage the unimodal subtasks to explore the modality-specific features. Furthermore, benefiting from the large scale of dataset, the knowledge transferring in distillation from teacher subtasks to student subtasks becomes more effective and smooth. The experiment results on CMU-MOSEI show that MTMD defeats MMIM by 0.6% on Acc7, and outperforms Self-MM by 2.2%/0.9% on Acc2, 2.1%/0.7% on F1, 0.016 on MAE and 0.004 on Corr. The better performance on two public MSA datasets demonstrates the superiority of the proposed MTMD.

3) Ablation Study: We perform ablation studies on CMU-MOSI to verify the effectiveness of the designed momentum distillation approach and show the role of multiple subtasks in the learning process, as shown in Table II.

Ablation on Dominant Modality Distillation (DMD)

1) Role of response-based knowledge distillation: After removing $\mathcal{L}^u_{reg}$ or $\mathcal{L}^u_{cls}$, there are performance drops on all metrics both under two circumstances. Specifically, $\mathcal{L}^u_{reg}$ has a greater impact on the regression metrics such as MAE and Corr while $\mathcal{L}^u_{cls}$ affects the classification metrics such as Acc7, which

TABLE II
ABLATION STUDY OF MTMD ON CMU-MOSI DATASET. NOTE THAT "W/O MOM" DENOTES WITHOUT MOMENTUM DISTILLATION, "W/O HARDATT" DENOTES REMOVING HARD ATTENTION STRATEGY IN MFD, AND "NO DISTILL" DENOTES REMOVING THE KNOWLEDGE DISTILLATION AMONG DIFFERENT SUBTASKS

Description	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑
(1) DMD					
w/o $\mathcal{L}^u_{reg}$	45.7	82.4/84.3	82.4/84.3	0.732	0.788
w/o $\mathcal{L}^u_{cls}$	45.4	82.7/84.3	82.6/84.3	0.728	0.791
w/o $\mathcal{L}^u_{feat}$	45.9	82.6/84.6	82.5/84.6	0.727	0.792
w/o queue	46.9	83.5/85.5	83.5/85.5	0.716	0.793
w/o MoM	46.1	82.7/84.6	82.6/84.6	0.731	0.791
(2) MFD					
w/o $w_{t/a/v}$	45.0	82.2/83.8	82.2/83.9	0.737	0.790
w/o HardAtt	46.2	83.2/85.4	83.1/85.4	0.715	0.794
w/o $\mathcal{L}^u_{fusion}$	46.0	82.5/84.4	82.5/84.4	0.727	0.790
w/o MoM	44.8	81.5/84.2	81.7/84.1	0.733	0.785
w/o t subtask	17.9	55.0/53.1	52.5/51.2	1.420	0.234
w/o a subtask	46.2	83.2/85.0	83.1/85.0	0.718	0.794
w/o v subtask	46.5	82.9/84.8	82.9/84.8	0.723	0.789
No Distill	44.0	81.8/83.2	81.7/83.3	0.749	0.780
MTMD	**47.5**	**84.0/86.0**	**83.9/86.0**	**0.705**	**0.799**

indicates the various contributions of regression and classification subtasks on different metrics.

2) Role of feature-based knowledge distillation: Removing $\mathcal{L}^u_{feat}$ results in a clear drop especially on Acc2 and F1. In feature-based contrastive distillation, the contrastive learning is presented based on the sentiment intensity classes, contributing mainly to the inductive learning of the representations in the feature space. Without the textual feature knowledge as the teacher guidance, we remark that the learning of inferior modalities representation becomes more unstable, which is due to the tendency of overfitting on the inherent noises in the acoustic and visual modalities. Besides, we remove the queues to examine the effect of the unimodal representation queues and observe the performance slightly drops. We surmise that the representation queues can benefit the contrastive learning by providing more samples as sentiment-related supervision which further improves the distillation of feature-based knowledge.

3) Role of momentum distillation of DMD: Without momentum distillation, the performance decreases significantly which indicates that the absence of prediction and feature supervision from momentum unimodal model impairs MTMDś capability in capturing intra-modal dynamics and causes representation learning to neglect the modality-specific features.

Ablation on Multimodal Fusion Distillation (MFD)

1) Role of momentum fusion factors: In Table II, removing the momentum fusion factors $\{w_t, w_a, w_v\}$ greatly harms the performance of the approach. Considering the direct influence of the factors on the fusion stage of different modalities, we deduce that they not only affect the robustness of multimodal representations but also have indirect effect on the unimodal subtasks as the teacher guidance in knowledge distillation. The further analysis of respective unimodal factor will be depicted in Section V-E4.

2) Role of the hard samples attention strategy: Focusing on the hard samples in a batch, the hard attention strategy increases the discrimination ability of the approach, especially bringing a performance boost by 1.3% on Acc7.

3) Role of multimodal knowledge distillation: Removing the guidance of multimodal subtask by $\mathcal{L}^m_{fusion}$ is considered as reducing a sentiment-related view at the level of both representation and prediction in the learning of unimodal subtask, which impairs the performance similar to remove dominant modality distillation.

4) Role of momentum distillation of MFD: The momentum distillation plays an important role in MFD similarly as DMD, without which the performance drops considerably. After removing the momentum fusion factors and the knowledge distillation from the multimodal subtask, the approach lacks ability to guide the unimodal subtask to capture cross-modal dynamics and learn a robust multimodal representation.

Ablation on Multi-Task Learning architecture

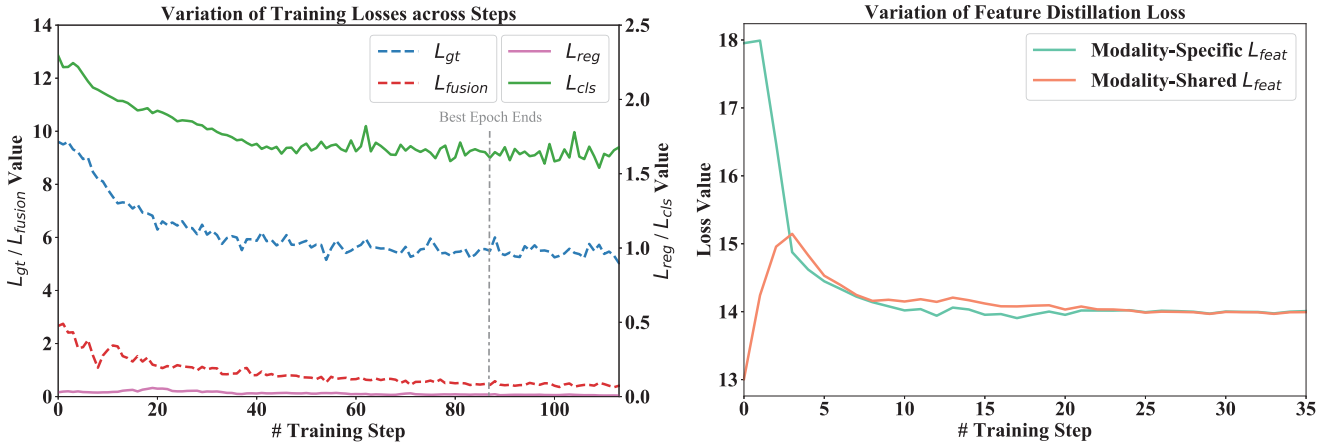
1) Roles of multiple subtasks: We observe that the performance drops sharply when the textual subtask is removed while drops relatively less when removing the other two unimodal subtasks. The ablation of textual, acoustic, and visual subtasks indicates that the textual modality stands in the significant dominant position over the other two inferior modalities in MSA, which confirms the same conclusion from prior researches [3], [11], [16], [75]. Moreover, these observations further reveal the reasonableness and effectiveness of the designed dominant modality distillation.

2) Role of distillation among different subtasks: Without knowledge distillation of the subtasks, the modality gaps existing among different modality limits the cross-modal interaction and the learning of modality-shared features. The modality-specific features extracted in multiple unimodal subtasks are difficult to model an efficient multimodal representation in the stage of fusion. Thus, the performance on all metrics is substantially weakened after removing the distillation.

E. Further Analysis

1) Losses Tracing: As shown in Fig. 5(a), we visualize the trends of the label-level losses including ground truth losses $\mathcal{L}_{gt}$, response-based momentum distillation losses $\mathcal{L}^m_{reg}$, $\mathcal{L}^m_{cls}$, $\mathcal{L}^m_{fusion}$. At the level of prediction, the ground truth losses $\mathcal{L}_{gt}$ smoothly descents benefited from the continuous interaction of intra- and inter-modal dynamics in the knowledge distillation. For the momentum distillation losses, we observe that there are step windows of rising losses in the optimization of $\mathcal{L}^m_{reg}$, $\mathcal{L}^m_{cls}$ and $\mathcal{L}^m_{fusion}$, which reveals the adversarial process of the ground truth hard labels with the soft pseudo labels generated by the momentum unimodal models as the guidance of unimodal subtasks. In general, the convergence of the losses at the label-level is stable and robust.

In the tracing of feature-level loss, we divide $\mathcal{L}^m_{feat}$ into modality-specific and -shared parts, and visualize the loss



(a) Label-level Losses $\mathcal{L}_{gt}$ / $\mathcal{L}^m_{reg}$ / $\mathcal{L}^m_{cls}$ / $\mathcal{L}^m_{fusion}$ (b) Modality-specific and -shared Feature-level Losses $\mathcal{L}^m_{feat}$

Fig. 5. Variation of losses during training on CMU-MOSI dataset. The values for plotting are the average losses in a constant interval of every 5 steps. Similar variations are also observed in CMU-MOSEI dataset.

variation as shown in Fig. 5(b). Specifically, the original feature contrastive losses $\mathcal{L}^u_{feat}$ for $u \in \{a, v\}$ are seen as the modality-shared feature distillation losses aiming at capturing the cross-modal dynamics across different modalities. While in the momentum distillation, the unimodal momentum model generates the pseudo-targets to further conduct contrastive learning with the original features encoded by the unimodal online model. The introduced momentum part in (11) is represented as modality-specific features contrastive loss for $u \in \{a, v\}$. In the learning process of feature-based knowledge distillation, due to the huge modality gap among different modalities, it's difficult to first reduce the modality-shared part in the whole $\mathcal{L}^m_{feat}$. Thus, MTMD tends to reduce the modality-specific $\mathcal{L}^m_{feat}$ at the beginning of training. Since capturing and retaining the most important modality-specific features, MTMD further explores the modality-shared features with the assistance of both response-based and feature-based knowledge distillation. The successful transferring of cross-modal information promotes the decrease of modality-shared $\mathcal{L}^m_{feat}$. Next, modality-specific and -shared $\mathcal{L}^m_{feat}$ are optimized in the meantime. Lastly, $\mathcal{L}^m_{feat}$ reaches the minimal value. Notably, the value of the modality-specific part eventually equals the one of the modality-shared part in the feature distillation loss, which illustrates that the modality-specific features are adaptively balanced with modality-shared features.

2) Contrastive Learning on Sentiment Class or Instance: In CRD [56], different negative sampling policies in contrastive learning bring different performance on the same dataset. When giving an anchor utterance input X_i, we apply two negative sampling policies in the contrastive learning of feature-based knowledge distillation: 1) X_j where $j \neq i$ for considering different modalities from the same utterance instance as the positive pairs while versa as the negative samples; 2) X_j with label $cls_j \neq cls_i$ for deeming the utterances in the same sentiment class as the positive pairs while versa as the negative samples, where $cls \in \{positive, neutral, negative\}$. As shown in Table III, feature-based contrastive distillation based on the sentiment classes performs better than distillation based on the utterance instance.

TABLE III
ABLATION STUDY OF DIFFERENT SAMPLING POLICIES, WHERE "$i \neq j$" DENOTES DEEMING SAMPLES FROM DIFFERENT UTTERANCES AS THE NEGATIVE SAMPLES AND "$cls_i \neq cls_j$" DENOTES DEEMING SAMPLES IN DIFFERENT SENTIMENT CLASSES AS THE NEGATIVE SAMPLES

Sampling	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑
$i \neq j$	45.9	83.1/84.8	83.0/84.7	0.719	0.791
$cls_i \neq cls_j$	**47.5**	**84.0/86.0**	**83.9/86.0**	**0.705**	**0.799**

Specifically, the instance-based negative sampling strategy implies the tendency of increasing the intra-class variance since the utterances from the same sentiment class may be pushed apart, which impacts the classification of sentiment intensity, especially for the utterances with a neutral sentiment. On the contrary, the sentiment-based negative sampling strategy avoids the issues. Moreover, the second strategy further explores the sentiment information contained in the representations among different modalities.

3) Visualization in the Embedding Space: We present a visualization for the distribution of multimodal representation learned on the testing set of CMU-MOSI in the embedding space by projecting the representations into 2-dimensional feature points utilizing the t-SNE algorithm [77]. As shown in Fig. 6(b), before distillation, the multimodal representation tends to be more dispersed in the embedding space and samples from different sentiment classes can not be classified accurately. Nevertheless, after the proposed momentum distillation approach, the multimodal representation is more distinguishable on the sentiment intensity. Besides, the representations from the same sentiment class are clustered compactly while the ones from diverse sentiment classes are pushed apart. Furthermore and interestingly, we observe that the representation points from positive and negative classes are separated in a linear variation way as shown in Fig. 6(b), which explicitly illustrates the sentiment labels annotated linearly ranging from −3 to +3. The multimodal

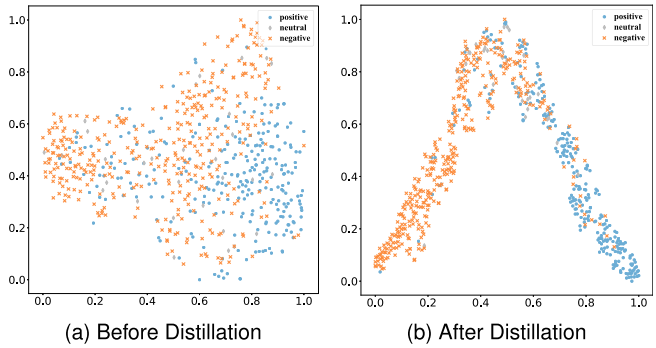


Fig. 6. T-SNE visualization of multimodal representation in the embedding space on the testing set of CMU-MOSI, where "blue dot", "gray diamond" and "orange x" denotes positive, neutral and negative sentiment, respectively.

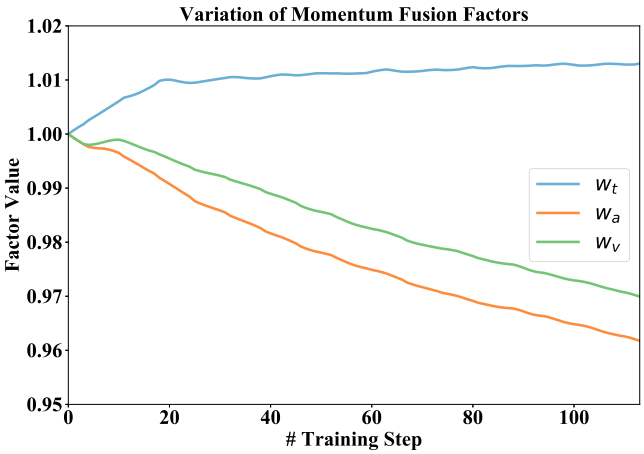


Fig. 7. Variation of the momentum fusion factors during training on the CMU-MOSI dataset.

representations learned by the knowledge distillation approach efficiently capture the sentiment-related features and sufficiently explore the intra- and inter-modal dynamics.

4) Tracing the Variation of Momentum Fusion Factors: As shown in Fig. 7, we initialize the momentum fusion factors $\{w_t, w_a, w_v\}$ with 1.0 and trace the variation of the factors during training on the CMU-MOSI dataset. It can be observed that the value of textual factor w_t is always higher than the values of acoustic and vision factor w_a, w_v, which strengthens the view that the textual modality plays a dominant role in the multimodal representation compared with the other two inferior modalities. Besides, the textual factor w_t increases and stabilizes soon, deducing that the model tends to extract the rich and robust semantic information contained in the textual representation. On the other hand, the acoustic and vision factors retain continuously declining during the training stage due to the inherent noises of the pre-processed acoustic and visual features. However, further reducing the weight of acoustic and vision representations in the multimodal representation after the best epoch deteriorates the performance, which indicates that the inferior modalities learned by knowledge distillation contribute to the final sentiment prediction. We infer the reason is that the modality-specific features of acoustic and visual modalities are productive in the correction of subjective and biased emotion issues of the textual modality, meaning that different modalities are complementary in the multimodal fusion.

VI. Conclusion

In this article, by regarding the learning process of different modalities as multiple subtasks, we propose a novel approach Multi-Task Momentum Distillation (MTMD) which adopts knowledge momentum distillation to transfer the sentiment-related information across different subtasks. Considering textual subtask and multimodal subtask as the teacher networks while acoustic and visual subtasks as the student ones, we guide the unimodal subtasks to concentrate on modality-shared features and efficiently capture the cross-modal dynamics. Specifically, we reduce the modality gap at the level of prediction and representation by response-based and feature-based knowledge distillation. Besides, we introduce the unimodal momentum model as an exponential-moving-average version of the unimodal online model to explore the modality-specific features in the knowledge distillation. Moreover, we present momentum fusion factors to adaptively adjust the contribution of different modalities in the multimodal fusion and finally obtain a robust multimodal representation to conduct the sentiment prediction task. In addition, we theoretically interpret the learning of MTMD as generating multiple sentiment-related views in a mutual information maximization perspective. Extensive experiments illustrate the effectiveness of our approach and the proposed MTMD achieves state-of-the-art performance on two public MSA datasets.

References

[1] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark: Assoc. Comput. Linguistics, 2017, pp. 1103–1114.

[2] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*. Florence, Italy: Assoc. Comput. Linguistics, 2019, pp. 6558–6569.

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

[4] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 108–132, First Quarter 2023.

[5] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Int. Conf. Mach. Learn.*, Madison, WI, USA: Omnipress, 2011, pp. 689–696.

[6] S. Katada, S. Okada, and K. Komatani, "Effects of physiological signals in different types of multimodal sentiment estimation," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2022.3155604.

[7] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proc. 13th Int. Conf. Multimodal Interfaces*, New York, NY, USA, 2011, pp. 169–176.

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

[9] Y. Wu, Z. Lin, Y. Zhao, B. Qin, and L.-N. Zhu, "A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis," in *Proc. Findings Assoc. Comput. Linguistics: ACL-IJCNLP*, 2021, pp. 4730–4738.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[11] W. Rahman et al., "Integrating multimodal information in large pretrained transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2359–2369.

[12] H. Mao, Z. Yuan, H. Xu, W. Yu, Y. Liu, and K. Gao, "M-SENA: An integrated platform for multimodal sentiment analysis," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics: Syst. Demonstrations*, Dublin, Ireland, 2022, pp. 204–213.

[13] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *Proc. Int. Conf. Learn. Representations*, 2019.

[14] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6892–6899.

[15] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 8992–8999.

[16] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 9180–9192.

[17] S. Mai, Y. Zeng, S. Zheng, and H. Hu, "Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2022.3172360.

[18] Y.-H. H. Tsai, M. Ma, M. Yang, R. Salakhutdinov, and L.-P. Morency, "Multimodal routing: Improving local and global interpretability of multimodal language analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 1823–1833.

[19] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 10 790–10 797.

[20] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[21] A. S. d'Avila Garcez and L. Lamb, "Neurosymbolic AI: The 3rd wave," *Artif. Intell. Rev.*, pp. 1–20, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s10462-023-10448-w

[22] A. N. Sheth, K. Roy, and M. Gaur, "Neurosymbolic AI - Why, what, and how," 2023, *arXiv:2305.00813*.

[23] E. Cambria, Q. Liu, S. Decherchi, F. Xing, and K. Kwok, "SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis," in *Proc. 13th Lang. Resour. Eval. Conf.*, Marseille, France: Eur. Lang. Resour. Assoc., 2022, pp. 3829–3839.

[24] N. Krishnaswamy and J. Pustejovsky, "Neurosymbolic AI for situated language understanding," 2020, *arXiv: 2012.02947*.

[25] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, "The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision," in *Proc. Int. Conf. Learn. Representations*, 2019.

[26] Q. Li, D. Gkoumas, C. Lioma, and M. Melucci, "Quantum-inspired multimodal fusion for video sentiment analysis," *Inf. Fusion*, vol. 65, pp. 58–71, 2021.

[27] L. Stappen, A. Baird, L. Schumann, and S. Bjorn, "The multimodal sentiment analysis in car reviews (MuSe-CaR) dataset: Collection, insights and improvements," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1334–1350, Second Quarter 2023.

[28] H. Wang, A. Meghawat, L.-P. Morency, and E. P. Xing, "Select-additive learning: Improving generalization in multimodal sentiment analysis," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2017, pp. 949–954.

[29] J. Williams, S. Kleinegesse, R. Comanescu, and O. Radu, "Recognizing emotions in video using multimodal DNN feature fusion," in *Proc. Grand Challenge Workshop Hum. Multimodal Lang.*, Melbourne, Australia, 2018, pp. 11–19.

[30] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.

[31] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5642–5649.

[32] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proc. IEEE 16th Int. Conf. Data Mining*, 2016, pp. 439–448.

[33] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 873–883.

[34] E. Cambria, D. Hazarika, S. Poria, A. Hussain, and R. Subramanyam, "Benchmarking multimodal sentiment analysis," in *Proc. Int. Conf. Comput. Linguistics Intell. Text Process.*, Springer, 2017, pp. 166–179.

[35] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5634–5641.

[36] J. Williams, O. Radu, R. Comanescu, and L. Tian, "DNN multimodal fusion techniques for predicting video sentiment," in *Proc. Grand Challenge Workshop Hum. Multimodal Lang.*, 2018, pp. 64–72.

[37] H. Pham, T. Manzini, P. P. Liang, and B. Poczós, "Seq2Seq2Sentiment: Multimodal sequence to sequence models for sentiment analysis," in *Proc. Grand Challenge Workshop Hum. Multimodal Lang.*, 2018, pp. 53–63.

[38] L. Hemamou, A. Guillon, J.-C. Martin, and C. Clavel, "Multimodal hierarchical attention neural network: Looking for candidates behaviour which impact recruiter's decision," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 969–985, Second Quarter 2023.

[39] J. Yu, K. Chen, and R. Xia, "Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2022.3171091.

[40] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 12, pp. 5586–5609, Dec. 2022.

[41] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, "Multi-task learning for dense prediction tasks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3614–3633, Jul. 2022.

[42] I. Kokkinos, "UberNet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5454–5463.

[43] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 525–536.

[44] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, "Cross-stitch networks for multi-task learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3994–4003.

[45] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1871–1880.

[46] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo, "Multi-task deep visual-semantic embedding for video thumbnail selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3707–3715.

[47] M. S. Akhtar, D. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multi-task learning for multi-modal emotion recognition and sentiment analysis," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 370–379.

[48] W. Yu et al., "CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3718–3727.

[49] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7482–7491.

[50] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *Proc. 35th Int. Conf. Mach. Learn.*, J. Dy and A. Krause, Eds., 2018, pp. 794–803.

[51] G. Hinton et al., "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[52] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, 2021.

[53] J. Kim, S. Park, and N. Kwak, "Paraphrasing complex network: Network compression via factor transfer," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA: Curran Associates Inc., 2018, pp. 2765–2774.

[54] N. Komodakis and S. Zagoruyko, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. Int. Conf. Learn. Representations*, 2017.

[55] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3962–3971.
[56] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," in *Proc. Int. Conf. Learn. Representations*, 2020.
[57] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2827–2836.
[58] L. Zhao, X. Peng, Y. Chen, M. Kapadia, and D. N. Metaxas, "Knowledge as priors: Cross-modal knowledge generalization for datasets without superior knowledge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6527–6536.
[59] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds. Curran Associates, Inc., 2021, pp. 9694–9705.
[60] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, M. Meila and T. Zhang, Eds., 2021, pp. 8748–8763.
[61] R. Zhang, Z. Zeng, Z. Guo, and Y. Li, "Can language understand depth?," in *Proc. 30th ACM Int. Conf. Multimedia*, New York, NY, USA, 2022, pp. 6868–6874.
[62] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[63] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[64] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735.
[65] P. Khosla et al., "Supervised contrastive learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020, pp. 18 661–18 673.
[66] D. B. F. Agakov, "The IM algorithm: A variational approach to information maximization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2004, Art. no. 201.
[67] A. V. D. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.
[68] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," 2016, *arXiv:1606.06259*.
[69] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*. Melbourne, Australia, 2018, pp. 2236–2246.
[70] P. Ekman, W. V. Freisen, and S. Ancoli, "Facial signs of emotional experience," *J. Pers. Social Psychol.*, vol. 39, no. 6, 1980, Art. no. 1125.
[71] P. P. Liang, Z. Liu, A. B. Zadeh, and L.-P. Morency, "Multimodal language analysis with recurrent multistage fusion," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, 2018, pp. 150–161.
[72] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proc. Conf. Assoc. Comput. Linguistics Meeting*, 2018, Art. no. 2225.
[73] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP — A collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 960–964.
[74] iMotions, "Facial expression analysis," 2017. [Online]. Available: https://imotions.com/
[75] D. Gkoumas, Q. Li, C. Lioma, Y. Yu, and D. Song, "What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis," *Inf. Fusion*, vol. 66, pp. 184–197, 2021.
[76] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*.
[77] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.