Hybrid Contrastive Learning of Tri-Modal Representation for Multimodal Sentiment Analysis

Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu

Abstract—The wide application of smart devices enables the availability of multimodal data, which can be utilized in many tasks. In the field of multimodal sentiment analysis, most previous works focus on exploring intra- and inter-modal interactions. However, training a network with cross-modal information (language, audio and visual) is still challenging due to the modality gap. Besides, while learning dynamics within each sample draws great attention, the learning of inter-sample and inter-class relationships is neglected. Moreover, the size of datasets limits the generalization ability of the models. To address the afore-mentioned issues, we propose a novel framework HyCon for hybrid contrastive learning of tri-modal representation. Specifically, we simultaneously perform intra-/inter-modal contrastive learning and semi-contrastive learning, with which the model can fully explore cross-modal interactions, learn inter-sample and inter-class relationships, and reduce the modality gap. Besides, refinement term and modality margin are introduced to enable a better learning of unimodal pairs. Moreover, we devise pair selection mechanism to identify and assign weights to the informative negative and positive pairs. HyCon can naturally generate many training pairs for better generalization and reduce the negative effect of limited datasets. Extensive experiments demonstrate that our method outperforms baselines on multimodal sentiment analysis and emotion recognition.

Index Terms—Multimodal sentiment analysis, supervised contrastive learning, representation learning, multimodal learning

1 Introduction

Thanks to the wide applications of smart devices, we can take advantage of abundant multimodal data to perform many downstream tasks [1]. As one of the main research directions of multimodal machine learning, multimodal sentiment analysis (MSA) aims to predict the sentiment score from audio, visual, and language features (see Fig. 1). MSA draws increasing attention with the availability of multiple sources of information, and richer information should help to boost performance. However, it has been a challenge to learn meaningful representations for multimodal data due to the modality gap. Researchers endeavor to devise effective models including RNN variants [2], [3], [4], [5], [6], [7], [8], Transformers [9], [10], [11], BERT-based models [12], [13], [14] to eliminate the gap and learn sufficient interactions between modalities. Those methods mostly focus on learning joint representations in a common manifold, using fusion methods to obtain cross-modal interactions for better performance.

However, most previous works cannot ensure to align modalities well with each other and still suffer modality gap [13], [16]. Those works mostly focus on learning interactions on an intra-class setting, or more specifically, intra-sample setting, which neglect the inter-sample and inter-class relationships. In other words, prior works seek to learn intra- and inter-modal dynamics within each sample, but not between samples, especially between those samples belonging to different classes. Besides, there exists a tendency of overfitting due to the small size of public datasets and the large amount of parameters introduced by sophisticated designed unimodal and fusion networks, which degrades the generalization performance of existing MSA models.

We hypothesize that regularizing the distance between different classes and different samples to explore the inter-class and inter-sample relationships help improve the quality of joint cross-modal embedding space, which will be more discriminative for classification/regression and of better generalization ability. A possible solution to implement this hypothesis is to utilize contrastive learning [17], which allows the model to sufficiently learn intra-/inter-class cross-modal dynamics. In recent years, a resurgence of work in contrastive learning has led to major advances in representation learning [17], [18], [19], [20]. Based on it, we propose a novel framework HyCon for hybrid contrastive learning of tri-modal representation, generating abundant positive and negative pairs of unimodal representations for supervised and unsupervised learning.

The technical novelty of HyCon is to elaborately devise novel losses based on latest contrastive learning literature

- *Sijie Mai, Ying Zeng, and Shuangjia Zheng are with Sun Yat-sen University, Guangzhou 510275, China. E-mail: {maisj, zengy268, zhengshj9}@mail2.sysu.edu.cn.*
- *Haifeng Hu is with School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510275, China. E-mail: huhaif@mail.sysu.edu.cn.*

Manuscript received 9 Jan. 2022; accepted 29 Apr. 2022. Date of publication 3 May 2022; date of current version 13 Sept. 2023.
This work was supported in part by the National Key R&D Program of China under Grant 2018YFB1601100, and in part by the National Natural Science Foundation of China under Grant 62076262.
(Corresponding author: Haifeng Hu.)
Recommended for acceptance by C. Lee.
Digital Object Identifier no. 10.1109/TAFFC.2022.3172360

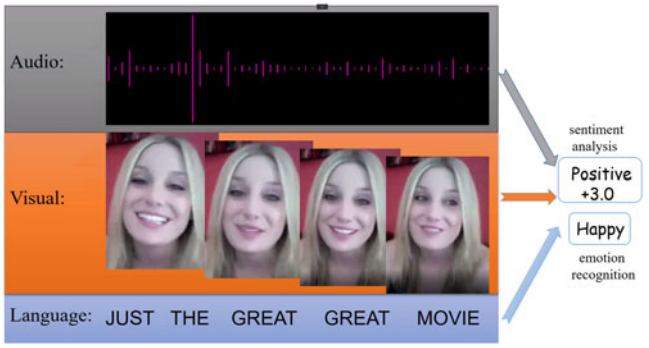


Fig. 1. A visualization of learning tasks. The sample comes from CMU-MOSI dataset [15]. The main task is multimodal sentiment analysis, and multimodal emotion recognition is used to evaluate the robustness of the model.

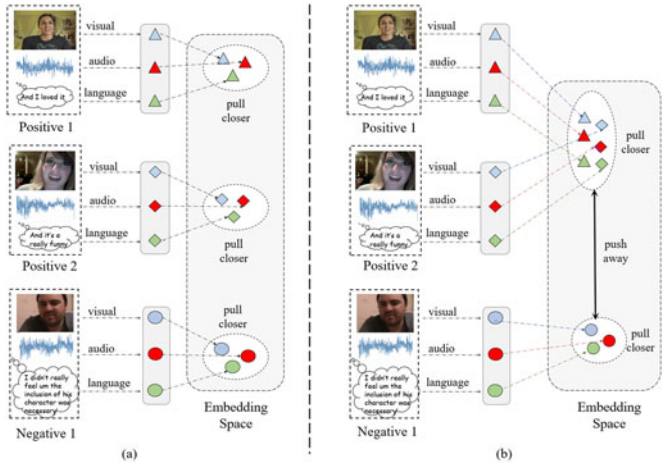


Fig. 2. A visualization of differences between the previous methods and our HyCon. The left sub-figure illustrates the effect of previous ideas [13], [16] that focus on learning the interactions and aligning modalities within each sample. The right one illustrates the key idea of HyCon: exploring inter-sample and inter-class relationships within and between modalities to learn a more discriminative embedding space and better align different modalities.

[19]. HyCon consists of three ways of contrastive learning, i.e., intra-modal contrastive learning (IAMCL), inter-modal contrastive learning (IEMCL) and semi-contrastive learning (SCL), to learn a discriminative cross-modal latent embedding. Specifically, addressing the issue that prior works focus on conserving the inter-modal interactions within each sample [13], [16], IAMCL and IEMCL allow HyCon to learn intra- and inter-modal dynamics between different samples in a supervised manner respectively (which is different from the traditional self-supervised contrastive learning). As shown in Fig. 2, by leveraging label information to learn the similarity of the pairs of unimodal representations and align different modalities from various samples, IAMCL and IEMCL ensure that unimodal representations from the same class are pulled closer than those from different classes in the feature space. Compared with IAMCL and IEMCL that mainly aim to learn a more discriminative embedding space, SCL aims to learn interactions between different modalities within each sample, which can be more focused on aggregating the distribution of different modalities and thus reduce the modality gap more effectively. Note that SCL only considers positive pairs of unimodal representations (that is why we call it semi-contrastive learning) to aggregate the modality distribution of the same sample. Moreover, a refinement term and the modality margin are introduced to enable a better representation learning capability. Specifically, the refinement term strengthens the learning of positive pairs to better align the unimodal representations from different samples, addressing the issue that positive pairs might not be sufficiently learned in previous contrastive learning methods [17], [20]. And instead of mapping all the unimodal representations to be the same [13], [16], [20], the modality margin allows the modality-specific information to be preserved when aligning different modalities, which is beneficial for fusion. These simple practices are shown to be effective in our experiments.

Training the network with redundant training pairs can be difficult. For representation learning, hard negative pairs are intuitively those pairs that are mapped nearby but should be far apart. Previous works have demonstrate that the hard negative pairs can help guide a model to correct its mistakes more quickly [21], [22]. Motivated by these works, to reduce the computational overhead and enable a more efficient learning, we employ a pair selection mechanism to dynamically select and highlight the informative training pairs. Note that previous works concentrate on the selection of hard negatives, mostly due to the reason that their positive samples are generated by augmentation instead of real [21], [22], [23], [24]. In contrast, we propose to identify both hard negative and hard positive pairs, and weight these pairs according to their hardness instead of treating these pairs equally.

As for the problem of limited size of datasets, HyCon takes advantage of contrastive learning to generate a large number of positive and negative pairs for training, with which HyCon can learn a more discriminative boundary between different classes in the feature space. The inherent advantage of contrastive learning helps to minimize the negative impact of limited datasets and reduce the possibility of overfitting, improving the representation learning ability of the proposed model.

In brief, the main contributions of this paper can be summarized as:

- *A novel learning framework for MSA:* A novel framework is proposed to learn tri-modal representation based on contrastive learning which naturally fits the nature of multimodal tasks. Considering inter-sample and inter-class relationships that are neglected in existing works, HyCon explores inter-class and inter-sample relationships to obtain a more discriminative joint embedding.
- *The proposed learning strategy:* Three losses, i.e., IAMCL, IEMCL and SCL, are devised to learn inter-/intra-modal dynamics in supervised and unsupervised manners comprehensively. To the best of our knowledge, we are the first to leverage contrastive learning in a hybrid manner to learn cross-modal embeddings. Instead of introducing complex module to align different modalities [13], [16], the devised losses introduce no additional parameters.
- *Extra designs for better optimization:* We devise the refinement term and the modality margin to enable a

better representation learning capability. The refinement term focuses on the learning of positive pairs to better align the information from different samples. And the modality margin allows the modality-specific information to be preserved when aligning modalities.

- *The design of the pair selection mechanism:* To improve the learning efficiency and focus on informative training pairs, we devise a pair selection mechanism to identify and highlight both the informative positive and negative pairs.

2 Related Work

2.1 Multimodal Sentiment Analysis

In recent years, multimodal sentiment analysis (MSA) has attracted significant research interest with the availability of multimodal data [1], [13], [25], [26]. Previous work mostly focus on designing sophisticated fusion strategies to explore inter-modal dynamics[27], [28].

Two direct ways to perform fusion are *early fusion* [29], [30], [31], [32] and *late fusion* [15], [33], [34], [35], [36]. Early fusion mainly concatenates the unimodal features at input level to conduct fusion, and late fusion tends to weighted average the unimodal decisions. Those methods outperform unimodal methods, but they can not fully explore intra-/inter-modal dynamics. With the continuous deepening of more advanced fusion methods, *tensor-based fusion* draws increasing attention for their high expressive power in exploring cross-modal dynamics [5], [37], [38]. Specifically, tensor fusion network (TFN) [37] and Low-rank Modality Fusion (LMF) [39] adopt outer product to learn both intra-modality and inter-modality dynamics end-to-end. Furthermore, some *modality translation methods* [10], [40] aim at learning a joint representation by translating source modality into the target one. Also, *graph fusion* is considered in [16], [41], who fuse the modalities using a graph fusion network and regard each interaction as one node. More recently, *BERT-based methods* such as Multimodal Adaptative Gate BERT (MAG-BERT) [12] and Cross-Modal BERT (CM-BERT) [14] achieve significant boost in performance by fine-tuning BERT with cross-modal information.

However, those works are heavily dependent on sophisticated fusion strategies, which introduce more parameters and higher computational costs. Besides, compared to unimodal methods, multimodal methods with complex fusion strategies may easily suffer from overfitting due to the significantly increased number of parameters. The overfitting problem is even more severe when a limited amount of training data is available. More importantly, those methods mostly focus on exploring intra-/inter-modal dynamics within each sample to narrow the gap between modalities, but neglect to consider inter-class relationships, which are also informative to help reach more discriminative boundaries and narrow down the modality gap.

Different from them, we investigate the effectiveness of contrastive learning to design a framework HyCon, which can avoid afore-mentioned problems. Our Hycon performs a combination of various ways of contrastive learning with very direct fusion strategy. Despite the simplicity of the fusion method, our proposed model still obtains new state-of-the-art results. To sum up, in addition to considering inter-class relationships for learning more sufficient intra-/inter-modal dynamics, our contrastive learning-based model has the inherent advantage to minimize the impact of limited dataset so as to learn generalized discriminative features. And it introduces no additional parameters for training, which reduces the probability of overfitting.

2.2 Contrastive Learning

In recent years, a resurgence of work in contrastive learning has led to major advances in many fields for its effectiveness in learning mutual information of different views of the data [17], [22], [42], [43]. Its principle is quite clear, following the idea that an anchor and a positive sample should be pulled closer in the feature space, while the anchor and negative samples should be pushed apart.

Traditional self-supervised contrastive learning requires one positive sample (e.g., an augmented version of the anchor sample) and many negative samples (e.g., randomly chosen samples from the mini-batch). Self-supervised contrastive learning methods try to learn features from different views, so as to distinguish samples from other samples [17], [18], [42]. But they only consider one positive sample in each mini-batch and cannot leverage label information even when it is available. Unlike those methods, SupCon [19] extends the self-supervised batch contrastive approach to the fully-supervised setting, and designs supervised contrastive loss (SupCon loss). Its positive samples are drawn from samples of the same class as the anchor, rather than being data augmentations of the anchor. SupCon loss can leverage label information, and the use of many positives and many negatives for each anchor allows us to achieve state-of-the-art performance without the need for hard-negative mining, which can be difficult to tune properly. However, we find that the learning with SupCon loss may fall into a sub-optimal solution when negative pairs are not sufficient.

Previous works based on contrastive learning mostly only has unimodal data (i.e., images). Some prior works have applied contrastive learning for cross-modal learning [44], [45]. Contrastive Bi-Modal Representation Algorithm (COBRA) [20] is the most related to our work, which represents the data across different modalities in a common manifold using contrastive learning paradigms. It achieves this goal by performing supervised contrastive learning, where a positive pair is defined as the representations of data samples belonging to the same modality and class, and a negative pair is defined as the representations of two samples belonging to same or different modalities of different classes. But COBRA does not utilize many positive samples in each mini-batch, and neglects the inter-modal interactions between different samples from the same class. Thus it may miss out some cross-modal dynamics.

Inspired by the previous works, we propose a novel framework HyCon for multimodal learning. With our hybrid contrastive learning strategy, HyCon can utilize label information for supervised contrastive learning, sufficiently learning cross-modal dynamics while at the same time exploring inter-class relationships. Moreover, we consider how to prevent the model from falling into a sub-optimal solution with a refinement term, and consider preserve

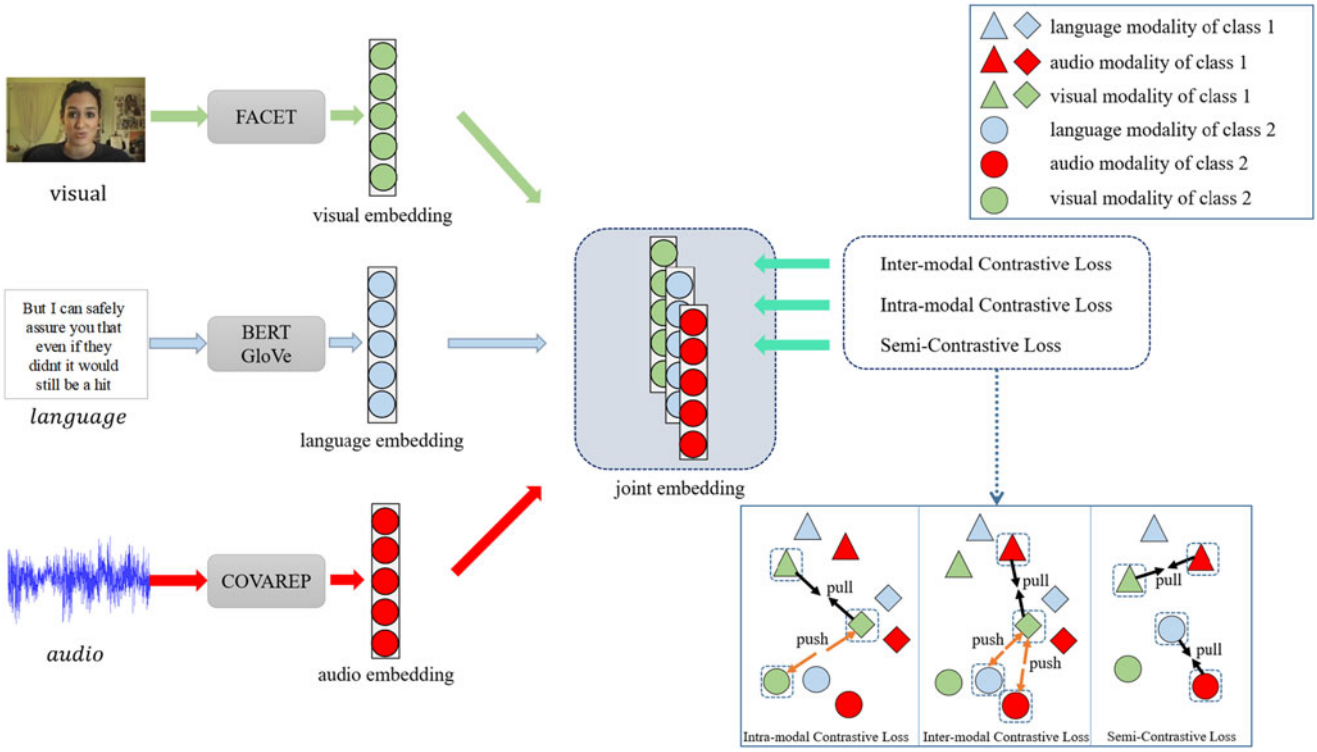


Fig. 3. The diagram of our proposed HyCon and the visualizations of the working of our designed losses.

the modality-specific information when aligning different modalities via the modality margin. In a word, our model can simultaneously preserve intra-/inter-modal interactions and inter-class relationships by constrastive learning in a hybrid manner.

3 Algorithm

In this section, we describe the proposed method in detail. We first explain the notations and the formulation of the problem statement, after which we introduce the pipeline of the proposed HyCon. The designed loss functions based on contrastive learning and the pair selection mechanism are then presented with emphasis. Finally, the optimization and training strategy will be explained.

3.1 Notations and Problem Formulation

Our task is to perform multimodal sentiment analysis with multimodal data by predicting the sentiment intensity or emotion score. The input to the model is an utterance [46] (i.e., a segment of a video bounded by pauses and breaths), each of which has three modalities, i.e., audio (a), visual (v), and language (l). The sequences of acoustic, visual, and language modalities are denoted as $\boldsymbol{u^a} \in \mathbb{R}^{d_a \times T_a}$, $\boldsymbol{u^v} \in \mathbb{R}^{d_v \times T_v}$, and $\boldsymbol{u^l} \in \mathbb{R}^{d_l \times T_l}$, where T_a, T_v and T_l represent the length of the audio, visual and language modality, respectively, and d_a, d_v and d_l denote the dimensionality of the audio, visual and language modality, respectively.

3.2 Overall Framework

In this section, we describe the pipeline of our proposed method, with its diagram illustrated in Fig. 3. Previous works mostly focus on designing fusion strategies to learn cross-modal interactions. However, they only consider intra-class (intra-sample) interactions but neglect the relationships of different samples, especially samples from different classes. Besides, the elaborately designed fusion strategies may introduce high computational costs and may lead to overfitting due to limited size of dataset. To this end, innovated by contrastive learning, we propose a novel framework HyCon to address MSA, which simultaneously considers intra-/inter-modal interactions and intra-/inter-class relationships. HyCon also achieves better representation learning ability with sufficient training pairs.

As shown in Fig. 3, given an input utterance, we first obtain the unimodal representations via unimodal learning networks. We then generate positive and negative pairs of unimodal representations according to different contrastive losses. After that, in the training stage, the model is trained with our designed contrastive learning to learn cross-modal interactions and inter-class relationships, which will be discarded at the inference time. Finally, the learned unimodal embeddings are fused with simple fusion method. To sum up, HyCon consists of the following main components:

- *Unimodal Learning Network*. For multimodal sentiment analysis task, to obtain the unimodal latent space of the three modalities, we leverage Transformer[9] to extract the features of the audio and visual modalities, and BERT[47] to process the language modality of the input utterance, respectively. For multimodal emotion recognition task, Transformer[9] is used to extract the language feature. The procedure can be formulated as:

$$\begin{aligned} \boldsymbol{X}^m &= \boldsymbol{F}^m(\boldsymbol{u}^m; \theta^m), m \in \{l, a, v\} \\ \boldsymbol{x}^m &= \boldsymbol{X}^m_{T_m} \in \mathbb{R}^{d \times 1}, \end{aligned} \tag{1}$$

where $\boldsymbol{F}^m$ parametrized by θ^m refers to each unimodal learning network, $\boldsymbol{u}^m \in \mathbb{R}^{d_m \times T_m}$ is the input sequence of modality m, and $\boldsymbol{x}^m$ is the projected unimodal representation. Note that $\boldsymbol{x}^m$ is the feature embedding of $\boldsymbol{X}^m$ in the last time step, and we only use the feature embedding in the last time step to conduct fusion such that our model is suitable for handling unimodal sequences of various length. The feature dimensionality d of unimodal representations is the same across all modalities.

- *Pair Generation*. We construct each mini-batch with K samples (each sample consists of audio, visual, language modalities). During the training stage, the positive pairs and negative pairs are randomly drawn from the mini-batch to train the model according to different losses. Pairs that are several times the number of samples can be generated, so that the model can maximize the use of datasets for better generalization ability.
- *Hybrid Contrastive Learning*. As the core of our proposed method, three versions of contrastive losses operated on the encoded unimodal representations are designed to perform intra-/inter-modal learning in the training stage. With the designed losses, HyCon can sufficiently learn dynamics within and between modalities, explore inter-class relationships and minimize the modality gap. The designed ways of contrastive learning are as follows:
  1) Semi-Contrastive Learning (SCL): SCL only considers positive pairs to learn interactions between different modalities of the same sample so as to minimize the modality gap.
  2) Intra-modal Contrastive Learning (IAMCL): IAMCL is performed in a supervised manner to learn intra-modal dynamics between different samples, considering multiple positive pairs and negative pairs in a mini-batch.
  3) Inter-modal Contrastive Learning (IEMCL): IEMCL is also performed in a supervised manner to learn inter-modal dynamics, and both IAMCL and IEMCL explore inter-class relationships.

  The designed contrastive learning will be explained in detail in Section 3.3.
- *Pair selection mechanism*. Pair selection mechanism is employ in HyCon-v2 to identify and highlight the informative positive and negative training pairs.
- *Fusion and Prediction*. With our designed contrastive learning strategy, HyCon can leverage simple fusion method to reach the state-of-the-art performance. In our work, different fusion strategies (including simple ways like concatenation) are adopted to demonstrate the effectiveness and generalization ability of HyCon, which is shown in Section 4.5.4. The defaulted fusion method is 'Element-wise Addition'. Generally, the fusion and prediction of our multimodal architecture can be summarized as

$$y_M = \boldsymbol{F}^M(\boldsymbol{x}^l, \boldsymbol{x}^a, \boldsymbol{x}^v; \theta_M), \ell = |y - y_M|, \tag{2}$$

where $\boldsymbol{F}^M$ parametrized by θ^M is the multimodal network, y is the ground truth label, y_M is the predicted label, and ℓ is mean absolute error (MAE) for prediction. Notably, for emotion recognition task, ℓ is defined as the cross-entropy loss instead of MAE.

3.3 HyCon

The traditional way to perform contrastive learning is self-supervised contrastive learning. A batch of randomly sampled pairs for each anchor consists of a positive pair and M negative pairs. The positive sample is an augmented version of the anchor sample, while each negative sample is an augmented version of the other samples. For each anchor sample, a simple self-supervised contrastive loss can be formulated as

$$\boldsymbol{L}_{self} = -log\frac{exp(\boldsymbol{a}^T\boldsymbol{p})}{exp(\boldsymbol{a}^T\boldsymbol{p}) + \sum_{j=1}^{M} exp(\boldsymbol{a}^T\boldsymbol{n}_j)}, \tag{3}$$

where $\boldsymbol{a}$, $\boldsymbol{p}$ and $\boldsymbol{n}$ denote the representation of the anchor, positive and negative sample, respectively.

To leverage label information for better classification performance, SupCon [19] extends the traditional self-supervised contrastive learning to the fully-supervised setting. In this setting, each batch is made up of N positive pairs and M negative pairs. A positive sample is any augmented version of samples from the same class, while each negative one comes from different classes. The use of many positives and many negatives for each anchor allows the model to achieve state-of-the-art performance without the need of hard-negative mining. For each anchor sample, supervised contrastive loss can be formulated as

$$\boldsymbol{L}_{sup} = -log\frac{\sum_{i=1}^{N} exp(\boldsymbol{a}^T\boldsymbol{p}_i)}{\sum_{i=1}^{N} exp(\boldsymbol{a}^T\boldsymbol{p}_i) + \sum_{j=1}^{M} exp(\boldsymbol{a}^T\boldsymbol{n}_j)}. \tag{4}$$

To apply contrastive learning in multimodal data, COBRA [20] adopts similar supervised contrastive loss to learn joint cross-modal embedding. The label information is leveraged in COBRA. For each anchor, the positive sample comes from the same class and the same modality, and the negative sample comes from the same or different modalities of different classes. The equation can be formulated as

$$\boldsymbol{L}_{COBRA} = -log\frac{\boldsymbol{a}^T\boldsymbol{p}}{\boldsymbol{a}^T\boldsymbol{p} + \sum_{j=1}^{M} \boldsymbol{a}^T\boldsymbol{n}_j}. \tag{5}$$

However, COBRA only considers one positive sample for each anchor as the way in traditional self-supervised contrastive learning. Besides, it only considers positive samples within the same modality but neglect those from different modalities. Consequently, the model may miss out some cross-modal dynamics, which hinders higher performance.

To sufficiently learn cross-modal dynamics and inter-class relationships, while at the same time fully leverages label information, we propose a novel framework HyCon to perform hybrid contrastive learnig in a supervised and unsupervised manner for MSA. In the training stage, hybrid contrastive learning is designed to train the model, which introduces three ways of contrastive learning and their corresponding contrastive losses. Different from previous contrastive learning methods, a refinement term is introduced

to enable a better representation learning for positive pairs and prevent the model from falling into a sub-optimal solution. Moreover, the modality margin is devised to preserve the modality-specific information when aligning different modalities, which is neglected in previous works [13], [16], [20]. We will explain them in detail as follows:

3.3.1 SCL: Semi-Contrastive Learning

We first introduce the simplest loss function, i.e., SCL, which is unsupervised. SCL is devised to learn inter-modal dynamics within the same utterance (sample), so as to minimize the modality gap. It only considers positive pairs (that is why we call it semi-contrastive learning), where the positive sample is defined as the unimodal representations from different modalities of the same utterance. For the encoded representation of each modality in each mini-batch $\boldsymbol{a}^m$ (i.e., $\boldsymbol{x}^m$), it generates two positive samples $S = \{\boldsymbol{p}_1^{m_1}, \boldsymbol{p}_2^{m_2}\}$ ($m, m_1, m_2 \in \{l, a, v\}$, $m \neq m_1$, $m_1 \neq m_2$ and $m \neq m_2$). The scoring function of each pair is based on the common dot product of the representations output by the unimodal learning network. First, for each modality in each sample, we perform $softmax$-normalization on the representation such that the dot product of each pair is between 0 and 1. The semi-contrastive loss can then be formulated as

$$\boldsymbol{L}_{SCL} = E_s\left[\frac{1}{2}\sum_{i=1}^{2}\left|\left|(\boldsymbol{a}^m)^T\boldsymbol{p}_i^{m_i} - \alpha\right|\right|^2\right],\ m \in \{l, a, v\}, \tag{6}$$

where E is an expectation operator over all the possible sets S in a mini-batch, and α is the modality margin between different modalities which allows for certain modality distributional discrepancies to retain the modality-specific information for fusion.

SCL allows the model to learn the dynamics between modalities of the same sample and draw their embeddings closer in the feature space, in which way the modality gap can be reduced. At the same time, considering that different modalities carry different discriminative modality-specific information that should not be completely eliminated, we allow for certain gap by applying modality margin α. Previous methods tend to explicitly match (translate) the representations of different modalities [10], [13], [16], [20], which is rather unrealistic. We argue that different modalities contain discrepancy modality-specific information, it is undesirable and also very difficult to map the representations from different modalities to the same one, which may lead to the loss of unimodal information. Therefore, we let the multimodal fusion stage to learn a richer multimodal representation that fuses information from various modalities rather than directly matching the representations of different modalities. The simple practice of introducing the modality margin term is shown to be effective in our experiments.

3.3.2 IAMCL: Intra-Modal Contrastive Learning

IAMCL aims to learn intra-modal dynamics and inter-class relationships for more discriminative boundaries in the feature space in a supervised manner. In IAMCL, a positive pair is defined as the two unimodal representations from the same modality and the same class of two different samples, while a negative pair is defined as the unimodal representations from the same modality of two samples whose classes are different. Considering a mini-batch of size K, for each modality m of each anchor in a mini-batch, it generates a set $S = \{\boldsymbol{p}_1^m, \boldsymbol{p}_2^m, \ldots, \boldsymbol{p}_N^m, \boldsymbol{n}_1^m, \boldsymbol{n}_2^m, \ldots, \boldsymbol{n}_M^m\}$ ($m \in \{l, a, v\}$), which consists of N positive samples and M negative samples ($N + M = K - 1$). Note that the size of mini-batch is fixed, but the number of N and M is random (i.e., the number of positive/negative pairs in a mini-batch is unfixed). Similar to SCL, $softmax$-normalization on the representations is performed such that the similarity of each pair is between 0 and 1. Then the intra-modal contrastive loss can be formulated as

$$\boldsymbol{L}_{IAMCL} = -E_s\left[\frac{\sum_{i=1}^{N}(\boldsymbol{a}^m)^T\boldsymbol{p}_i^m}{\sum_{i=1}^{N}(\boldsymbol{a}^m)^T\boldsymbol{p}_i^m + \sum_{j=1}^{M}(\boldsymbol{a}^m)^T\boldsymbol{n}_j^m}\right], \tag{7}$$

where $m \in \{l, a, v\}$ denotes the modality m, $\boldsymbol{a}^m$ denotes the representation of the anchor, $\boldsymbol{p}_i^m$ and $\boldsymbol{n}_j^m$ represents the representation of positive sample and negative sample, respectively. Observing Eq. (7), we can notice that the positive and negative samples come from the same modality as the anchor, but their classes are different. Moreover, the learning with traditional contrastive loss as Eq. (7) will be likely to fall into a sub-optimal solution where $\boldsymbol{a}^T\boldsymbol{n}_j$ is minimized but $\boldsymbol{a}^T\boldsymbol{p}_j$ is not maximized. This is because when the similarity of the negative pair approximates zero, the value of the loss is nearly minimized, regardless of what the similarity of the positive pair is. This phenomena can be more serious when the number of negative pairs is rare, where $\sum_{j=1}^{M}(\boldsymbol{a}^m)^T\boldsymbol{n}_j^m$ can be easily learned to be close to 0. In our algorithm, we hope that $(\boldsymbol{a}^m)^T\boldsymbol{n}_j^m$ is minimized and $(\boldsymbol{a}^m)^T\boldsymbol{p}_j^m$ is maximized. Therefore, we introduce a 'refinement term' to ensure that the similarity of the positive pairs can be maximized

$$\boldsymbol{L}_{IAMCL}^R = E_s\left[\frac{1}{N}\sum_{i=1}^{N}\left|\left|(\boldsymbol{a}^m)^T\boldsymbol{p}_i^m - 1\right|\right|^2\right],\ m \in \{l, a, v\} \tag{8}$$

$$\boldsymbol{L}_{IAMCL} \leftarrow \boldsymbol{L}_{IAMCL} + \boldsymbol{L}_{IAMCL}^R, \tag{9}$$

where $\boldsymbol{L}_{IAMCL}^R$ is the refinement loss for IAMCL. IAMCL encourages the representations of the same modality from different samples but belong to the same class to have the highest similarity, and forces the representations of the same modality from different classes to have the lowest similarity. With IAMCL, intra-modal dynamics and inter-class relationships can be learned. With the use of many positive and negative pairs, intra-modal interactions between different samples can be sufficiently explored.

3.3.3 IEMCL: Inter-Modal Contrastive Learning

IEMCL is similar to IAMCL except that IEMCL aims to learn inter-modal dynamics via contrastive learning, which is neglected in COBRA [20]. Also, IEMCL is different from SCL in that SCL focuses on learning inter-modal interactions within the same sample, while IEMCL explores inter-modal interactions between different samples to better align modalities and learn a more discriminative embedding space. Specifically, in IEMCL, a positive pair is defined as the two unimodal representations with different modalities

from different samples of the same class, while a negative pair is defined as the unimodal representations with different modalities from two samples whose classes are different. So, for an anchor in a mini-batch of size K, compared to IAMCL, it has twice as many negative and positive pairs as the IAMCL. After the $softmax$-normalization on all the unimodal representations, the IEMCL loss can be formulated as

$$\boldsymbol{L}_{IEMCL} = -E_s\left[\frac{\sum_{i=1}^{2N}(\boldsymbol{a}^m)^T\boldsymbol{p}_i}{\sum_{i=1}^{2N}(\boldsymbol{a}^m)^T\boldsymbol{p}_i + \sum_{j=1}^{2M}(\boldsymbol{a}^m)^T\boldsymbol{n}_j}\right], \tag{10}$$

where $m \in \{l, a, v\}$ denotes the modality m, $\boldsymbol{p}_i$ and $\boldsymbol{n}_j$ does not share the same modality as $\boldsymbol{a}^m$. Moreover, similar to IAMCL loss, we introduce a 'refinement term' to further strengthen the learning of the similarity of the positive pairs. And similar to SCL, a modality margin is also introduced to retain the modality-specific information for fusion

$$\boldsymbol{L}^R_{IEMCL} = E_s\left[\frac{1}{2N}\sum_{i=1}^{2N}\left|\left|(\boldsymbol{a}^m)^T\boldsymbol{p}_i - \alpha\right|\right|^2\right],\ m \in \{l, a, v\} \tag{11}$$

$$\boldsymbol{L}_{IEMCL} \leftarrow \boldsymbol{L}_{IEMCL} + \boldsymbol{L}^R_{IEMCL}, \tag{12}$$

where $\boldsymbol{L}^R_{IEMCL}$ is the refinement loss for IEMCL.

3.4 HyCon-v2: Explore the Selection of Pairs

The number of the positive and negative pairs of L_{IEMCL} and L_{IAMCL} increases as the size of the mini-batch K becomes larger. When K is a large number, the number of training pairs can also be very large, which leads to considerable computational overhead. Moreover, a large portion of the training pairs might not be useful, which might overwrite the effect of the informative training pairs. It is well known that making the negatives 'hard' is beneficial for representation learning, helping the model to correct its mistakes more quickly and effectively [21], [22]. To this end, we employ a non-parametric pair selection mechanism to dynamically select and highlight the effective pairs for training and discard the uninformative pairs. Note that a large body of prior works focus on hard negative mining [21], [22], [23], [45], [48], while we demonstrate the importance of both hard positive mining and hard negative mining, and assign different weights to these hard pairs according to their hardness.

Specifically, we compute the dot product of each pair, which is referred as the similarity of the pair, and then select the positive pairs that have low similarity and the negative pairs that have high similarity for training. For anchor $\boldsymbol{a}$, the procedures are shown as follows:

$$\boldsymbol{s^p} = \underset{h}{\mathrm{Min}}(\boldsymbol{a}^T\boldsymbol{p}), \quad \boldsymbol{s^n} = \underset{h}{\mathrm{Max}}(\boldsymbol{a}^T\boldsymbol{n}), \tag{13}$$

where $\boldsymbol{p} \in \mathbb{R}^{d\times n_p}$ and $\boldsymbol{n} \in \mathbb{R}^{d\times n_n}$ are the set of positive samples and negative samples for anchor $\boldsymbol{a}$, respectively. The Min_h function denotes that the smallest h scores of $\boldsymbol{a}^T\boldsymbol{p}$ are taken as $\boldsymbol{s^p}$, and Max_h likewise explains. We select a total number of $h = K \times \beta$ positive and negative pairs for $\boldsymbol{L}_{IEMCL}$ and $\boldsymbol{L}_{IAMCL}$, respectively. The hyperparameter β is flexible and can be determined by the users. We then assign weights to the selective pairs according to their hardness

$$w_i^n = \frac{exp(s_i^n)}{\sum_{j=1}^{h} exp(s_j^n)}, \quad w_i^p = \frac{exp(\frac{1}{s_i^p+\gamma})}{\sum_{j=1}^{h} exp(\frac{1}{s_j^p+\gamma})} \tag{14}$$

$$\hat{s}_i^n = s_i^n \cdot w_i^n, \quad \hat{s}_i^p = s_i^p \cdot w_i^p, \tag{15}$$

where $\hat{s}_i^n$ and $\hat{s}_i^p$ are the weighted similarity score of negative pair i and positive pair i respectively, and γ is a positive scalar to ensure that $s_*^p + \gamma$ is larger than zero. Note that we apply a gradient truncation technique to stop the gradient from the weight w_i^*, such that it only serves as a linear scalar to assist the training instead of learnable. By assigning a relatively high weight to the positive pairs that have low similarity and the negative pairs that have high similarity, the model can be learned more quickly and effectively. Interestingly, previous work [23] discovered that using the hardest negative pairs are harmful to the overall performance, which however is inapplicable to our problem. This is because the hardest negative examples in [23] are the examples that share the same class as the anchor, and we do not have such examples because the proposed $\boldsymbol{L}_{IEMCL}$ and $\boldsymbol{L}_{IAMCL}$ are supervised. We refer the HyCon version that employs the pair selection mechanism as HyCon-v2. The difference between the HyCon and the HyCon-v2 only lies in the selection of training pairs. We compare the performance of the HyCon and the HyCon-v2 in the experiment section to investigate the importance of pair selection.

3.5 Training

The overall contrastive loss function is a weighted sum of IAMCL, IEMCL and SCL, which can be formulated as

$$\boldsymbol{L}_{hybrid} = \lambda_1 L_{IAMCL} + \lambda_2 L_{IEMCL} + \lambda_3 L_{SCL}, \tag{16}$$

where λ_1, λ_2 and λ_3 are hyperparameters to constrain the contributions of the three contrastive losses. L_{hybrid} is summed over all data samples of different modalities in a mini-batch. The overall contrastive loss together with the prediction loss (see Eq. (2)) constitute the loss function for our model

$$\boldsymbol{L}_{overall} = \ell + \boldsymbol{L}_{hybrid}, \tag{17}$$

where $\boldsymbol{L}_{overall}$ is the overall loss function, and ℓ denotes the predictive loss.

4 Experiment

In the experiment section, we mainly evaluate our model on the multimodal sentiment analysis task. To evaluate the generalization ability of our model to other multimodal learning tasks, we additionally present the results of our model on the multimodal emotion recognition task. The detailed experimental setting and results are presented as follows.

4.1 Datasets

4.1.1 CMU-MOSI

CMU-MOSI[15] is a widely-used dataset for multimodal sentiment analysis, which consists of a collection of 2199 opinion video utterances downloaded from websites. Each opinion video is annotated with a sentiment intensity from

on -3 to +3. Following previous works [10], [12], we utilize 1,284 utterances for training, 229 utterances for validation, and 686 utterances for testing.

4.1.2 CMU-MOSEI

CMU-MOSEI[41] is a large dataset of multimodal sentiment analysis and emotion recognition. The dataset consists of over 20k video utterances from more than 1,000 YouTube speakers, covering 250 distinct topics. All the sentences utterance are randomly chosen from various topics and monologue videos, and each utterance is annotated on two views: emotion of six different values, and sentiment in the range [-3,3]. In our work, we use the sentiment label to perform MSA. We use 16,265 utterances for training, 1,869 utterances for validation, and 4,643 utterances for testing, which is the same as the previous works [10], [12].

4.1.3 IEMOCAP

IEMOCAP [49] is a multimodal emotion recognition dataset that contains a total number of 151 videos from 10 speakers. The videos are segmented into about 10K utterances. The dataset has the following labels: anger, happiness, sadness, neutral, excitement, frustration, fear, surprise and other. We take the first four emotions to compare with our baselines. We follow previous works [10], [50] to report the classification accuracy and the F1 score of each emotion.

4.2 Evaluation Metrics

The evaluation metrics for CMU-MOSEI and CMU-MOSI datasets are the same. We apply various metrics to evaluate the performance of each model: 1) Acc7: 7-way accuracy, sentiment score classification; 2) Acc2: binary accuracy, positive or negative; 3) F1 score; 4) MAE: mean absolute error (the lower the better); and 5) Corr: the correlation between the model's prediction and that of humans (the higher the better).

For the IEMOCAP dataset (multimodal emotion recognition task), we follow prior works [10], [50] to report the accuracy and F1 score of each emotion.

4.3 Baselines

1) *Early Fusion LSTM* (*EF-LSTM*), which concatenates the input features of different modalities at word-level, and then sends the concatenated features to an LSTM layer followed by a classifier to make prediction; 2) *Late Fusion LSTM* (*LF-LSTM*) uses an LSTM network for each modality to extract unimodal features and infer decision, and then combine the unimodal decisions by voting mechanism; 3) *Recurrent Attended Variation Embedding Network* (*RAVEN*) [51], which models interactions by shifting language representations based on the features of the audio and visual modalities; 4) *Memory Fusion Network* (*MFN*) [4], which proposes delta-attention module and multi-view gated memory network to discover inter-modal interactions; 5) *Multimodal Transformer* (*MULT*) [10], which learns multimodal representation by translating source modality into target modality using cross-modal Transformer [9]; 6) *Graph Fusion Network* (*GFN*) [16] uses a graph neural network to model unimodal, bimodal, and trimodal interactions. 7) *Tensor Fusion Network* (*TFN*) [37], which applies outer product from unimodal embeddings to jointly learn unimodal, bimodal and trimodal interactions; 8) *Low-rank Modality Fusion* (*LMF*) [39], which leverages low-rank weight tensors to reduce the complexity of tensor fusion; 9) *Quantum-inspired Multimodal Fusion* (*QMF*) [26], which focuses on the interpretable ability of multimodal fusion by taking inspiration from the quantum theory; 10) *Interaction Canonical Correlation Network* (*ICCN*) [52], which fuses with language embeddings with visual and audio features respectively to get two bimodal representations. Finally, the bimodal representations are fed to a Canonical Correlation Analysis (CCA) network to generate trimodal representation and make prediction; 11) *Multimodal Adaption Gate BERT* (*MAG-BERT*) [12], which proposes an attachment called Multimodal Adaptation Gate derived from RAVEN[51] that enables BERT [47] and XLNet [53] to accept multimodal data during fine-tuning. 12) *Hierarchical feature fusion network* (*HFFN*) [5] applies a hierarchical 'Divide, Conquer, and Combine' fusion strategy to fuse the three modalities for multimodal learning. 13) *Graph Memory Fusion Network* (*Graph-MFN*) [41] extends MFN [4] by a dynamic graph fusion to fuse the memory of the unimodal LSTMs. 14) *Temporal Convolutional Multimodal LSTM* (*TCM-LSTM*) [50] uses temporal convolutional network to extract the high-level unimodal representations and designs visual-/acoustic-LSTMs for multimodal fusion.

4.4 Experimental Details

(1) Baseline Evaluation. For each baseline on CMU-MOSI and CMU-MOSEI (multimodal sentiment analysis), following Gkoumas *et al.* [54], we first perform fifty-times random grid search on the hyper-parameters to fine-tune the model, and save the hyper-parameter setting that reaches the best performance. After that, we train each model with the best hyper-parameters setting for five times, and the final results are the mean results of the five-time running. Notably, for a fair comparison, for the baselines of multimodal sentiment analysis that use GloVe word embeddings [55] as the language embedding, we develop a counterpart of these baselines that use BERT [47] as the language encoder, which is the same as our method. Note that since the codes of QMF [26] and ICCN [52] are unavailable, we directly present the results from their papers. For the baselines on IEMOCAP (multimodal emotion recognition task), we borrow the results from [50].

(2) Feature Extraction. For CMU-MOSEI dataset, the input dimensionality of language, audio, and visual modality is 768, 74, and 35, respectively. For CMU-MOSI, the input dimensionality of language, audio, and visual modality is 768, 74, and 47, respectively. For IEMOCAP, the input dimensionality of the three modalities is 300, 74, and 35, respectively. For feature extraction, Facet[1] is used for the visual modality to extract a set of features that are composed of facial action units, facial landmarks, head pose, etc. These visual features are extracted from the utterance at the frequency of 30Hz to form a sequence of facial gestures over time. COVAREP [56] is utilized for extracting the representations of audio modality, which include 12 Mel-frequency cepstral coefficients, pitch tracking, speech polarity, glottal closure instants, spectral envelope, etc. These acoustic features are extracted from the full audio clip

TABLE 1
The Comparison With Baselines on CMU-MOSI

	Acc7	Acc2	F1	MAE	Corr
EF-LSTM-GloVe	31.6	75.8	75.6	1.053	0.613
EF-LSTM-BERT	43.9	82.0	82.1	0.771	0.791
LF-LSTM-GloVe	31.6	76.4	75.4	1.037	0.620
LF-LSTM-BERT	45.0	82.8	82.8	0.779	0.784
TFN-GloVe [37]	32.2	76.4	76.3	1.017	0.604
TFN-BERT [37]	44.7	82.6	82.6	0.761	0.789
LMF-GloVe [39]	30.6	73.8	73.7	1.026	0.602
LMF-BERT [39]	45.1	84.0	84.0	0.742	0.785
MFN-GloVe [4]	32.1	78.0	76.0	1.010	0.635
MFN-BERT [4]	44.1	83.5	83.5	0.759	0.786
MULT-GloVe [10]	33.6	79.3	78.3	1.009	0.667
MULT-BERT [10]	41.5	83.7	83.7	0.767	0.799
QMF-GloVe [26]	35.5	79.7	79.6	0.915	0.696
GFN-BERT [16]	47.0	84.3	84.3	0.736	0.790
ICCN-BERT [52]	39.0	83.0	83.0	0.860	0.710
RAVEN-GloVe [51]	33.8	78.8	76.9	0.968	0.667
MAG-BERT [12]	42.9	83.5	83.5	0.790	0.769
MAG-XLNet [12]	42.3	84.8	84.8	0.746	0.804
HyCon-BERT	46.6	85.2	85.1	0.713	0.790
HyCon-XLNet	46.0	85.5	85.4	0.688	0.818
HyCon-v2-BERT	47.7	85.7	85.6	0.705	0.805
HyCon-v2-XLNet	**48.3**	**86.4**	**86.4**	**0.664**	**0.832**

Note that the codes of QMF and ICCN is unavailable and we directly present the results from the original papers.

TABLE 2
The Comparison With Baselines on CMU-MOSEI Dataset

	Acc7	Acc2	F1	MAE	Corr
EF-LSTM-GloVe	46.7	79.1	78.8	0.665	0.621
EF-LSTM-BERT	52.0	84.4	84.5	0.608	0.781
LF-LSTM-GloVe	49.1	79.4	80.0	0.625	0.655
LF-LSTM-BERT	51.7	84.5	84.5	0.607	0.780
TFN-GloVe [37]	49.8	79.4	79.7	0.610	0.671
TFN-BERT [37]	51.8	84.5	84.5	0.622	0.781
LMF-GloVe [39]	50.0	80.6	81.0	0.608	0.677
LMF-BERT [39]	51.2	84.2	84.3	0.612	0.779
MFN-GloVe [4]	49.1	79.6	80.6	0.618	0.670
MFN-BERT [4]	52.6	84.8	84.8	0.607	0.771
MULT-GloVe [10]	48.2	80.2	80.5	0.638	0.659
MULT-BERT [10]	50.7	84.7	84.6	0.625	0.775
QMF-GloVe [26]	47.9	80.7	79.8	0.640	0.658
GFN-BERT [16]	51.8	85.0	85.0	0.611	0.774
ICCN-BERT [52]	51.6	84.2	84.2	**0.565**	0.713
RAVEN-GloVe [51]	50.2	79.0	79.4	0.605	0.680
MAG-BERT [12]	51.9	85.0	85.0	0.602	0.778
MAG-XLNet [12]	52.7	85.4	85.4	0.581	**0.797**
HyCon-BERT	52.8	85.4	85.6	0.601	0.776
HyCon-XLNet	53.2	86.4	**86.4**	0.590	0.788
HyCon-v2-BERT	52.5	85.8	85.7	0.590	0.784
HyCon-v2-XLNet	**53.4**	**86.5**	**86.4**	0.590	0.792

of each utterance at the sampling rate of 100Hz to form a sequence that represents variations in the tone of voice across the utterance. For CMU-MOSI and CMU-MOSEI, following the state-of-the-art methods [12], [13], [14], BERT [47] is used to extract the high-level language representation. Following the state-of-the-art methods in IEMOCAP [10], [50], GloVe word embeddings [55] are used to extract the features of the texts in the videos.[1]

(3) Training Details. We develop our model with the PyTorch framework on GTX1080Ti with CUDA 10.1 and torch version of 1.1.0. Our proposed model is trained using Adam [57] optimizer whose learning rate is set to 2e-5. The modality margin α is set to 0.8. For convenience, when generating the positive and negative pairs, we only consider the positive and negative classes of sentiment instead of 7-class fine-grained sentiment. λ_1, λ_2, and λ_3 are all set to 1. The feature dimensionality d is set to 50. The hyper-parameter β for the pair selection mechanism is set to 0.2.

4.5 Experimental Results

4.5.1 Evaluation on Multimodal Sentiment Analysis

In this section, we compare our proposed HyCon with other baselines on two datasets CMU-MOSI [58] and CMU-MOSEI [41]. As shown in Tables 1 and 2, although the methods that use BERT[47] as the language encoder (such as MAG-BERT[12] and GFN-BERT[16]) significantly outperform other existing methods that use GloVe[55] as the language encoder, it still can be seen that both the versions of our HyCon outperform these competitive baselines in most cases. Specifically, on CMU-MOSI dataset, our HyCon (the BERT version) outperforms the baselines in terms of Acc2, F1 score and MAE, and the HyCon-v2 outperforms baselines in terms of all the evaluation metrics. HyCon-v2 (the BERT version) yields 1.4% improvement on Acc2 and 1.3% improvement on F1 score compared to the best performing baseline GFN-BERT[16]. On CMU-MOSEI dataset, our HyCon (the BERT version) yields 0.9% improvement on Acc7, and 0.4% on Acc2 and 0.6% on F1 score compared to the state-of-the-art MAG-BERT, and the improvement of HyCon-v2 is more significant. These results demonstrate the effectiveness of our proposed model, indicating the importance of learning intra-/inter-modal dynamics and inter-class relationships in our HyCon. As for the comparison between the HyCon-v2 and HyCon, we can infer from Tables 1 and 2 that HyCon-v2 consistently outperforms HyCon on two datasets, suggesting the importance of selecting and highlighting the hard positive and negative samples.

For the sake of fair comparison, we also present the results of the baselines with BERT [47] as the language encoder. As shown in Tables 1 and 2, when GloVe is replaced by the BERT, there is a significant gain on the overall performance. For example, on the CMU-MOSI dataset, the Acc2 of the LMF-BERT[39] is about 10% higher than that of its GloVe counterpart. Nevertheless, our HyCon-v2 still outperforms the LMF-BERT by 1.7% in terms of Acc2. The improvement of our HyCon and HyCon-v2 is observed on both datasets compared to the baselines that use the same feature extraction method, which further proves the effectiveness of HyCon.

Moreover, it is shown in [12] that the XLNet[53] version of the model (MAG-XLNet) outperforms the BERT[47] version (MAG-BERT). Therefore, for more comprehensive comparison, we also present the results of our HyCon using XLNet as the text encoder. As presented in Tables 1 and 2, using XLNet brings consistent improvement on the overall

1. iMotions 2017. https://imotions.com/

TABLE 3
Comparison Between HyCon and Other Approaches on IEMOCAP Dataset

	Happy		Sad		Angry		Neutral		Average	
Models	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
MFN [4]	86.5	84.0	83.5	82.1	85.0	83.7	69.6	69.2	81.2	79.8
Graph-MFN [41]	86.8	84.2	83.8	83.0	85.8	85.5	69.4	68.9	81.5	80.4
RAVEN [51]	87.3	85.8	83.4	83.1	87.3	86.7	69.7	69.3	81.9	81.2
LMF [39]	86.9	82.3	85.4	84.7	87.1	86.8	**71.6**	**71.4**	82.8	81.3
MULT [10]	87.4	84.1	84.2	83.1	88.0	87.5	69.9	68.4	82.4	80.8
HFFN [5]	86.8	82.1	84.4	84.5	86.6	85.8	69.6	69.3	81.9	80.4
TCM-LSTM [50]	87.2	84.8	84.4	84.9	89.0	88.6	71.3	71.2	83.0	82.4
HyCon	**88.0**	85.5	86.2	85.9	89.4	89.2	70.4	70.5	83.5	82.8
HyCon-v2	87.5	**86.1**	**86.7**	**86.4**	**89.7**	**89.6**	71.2	71.0	**83.8**	**83.3**

performance. Moreover, the HyCon-XLNet and HyCon-v2-XLNet still outperforms the MAG-XLNet, which demonstrates the superiority of our contrastive learning strategy. Generally, these results suggest that using more advanced text encoder brings considerable improvement to the multimodal sentiment analysis systems, and our methods consistently outperform baselines under different text encoders.

4.5.2 Evaluation on Multimodal Emotion Recognition

We additionally evaluate the proposed methods on the task of multimodal emotion recognition to justify the generalization ability of the model to other multimodal learning task. The widely-used dataset IEMOCAP are evaluated in this section. Note that following the state-of-the-art methods [10], [50], we use GloVe[55] as the feature extractor of the language modality. From the results presented in Table 3, it can be seen that the original HyCon outperforms the baselines in the task of recognizing the 'Angry' and 'Sad' emotions, and it also performs better than the baselines in terms of the accuracy of recognizing the 'Happy' emotion. Moreover, the HyCon-v2 even outperforms HyCon and reaches the best performance of recognizing the 'Angry' and 'Sad' emotions, yielding about 1.5% improvement compared with the best results of baselines on the recognizing of the 'Sad' emotion. HyCon-v2 also achieves the best performance on the F1 score of recognition the 'Happy' emotion. Notably, HyCon and HyCon-v2 both outperform the baselines in terms of the average performance. In addition to the task of multimodal sentiment analysis, the extra experiments of multimodal emotion recognition have proven the effectiveness and generalization ability of the proposed methods.

4.5.3 Ablation Study

In this section, we perform ablation studies to verify the effectiveness of the designed contrastive learning method. And we further investigate the contribution of each component in thre contrative losses and the pair selection mechanism by removing it from the model.

Aiming to verify the effectiveness of the designed contrastive losses, we perform ablation experiment where all *contrastive losses* are removed (see the case of 'W/O Contrast Learning' in Table 4). From the experimental result, it can be seen that removing the whole contrastive learning method significantly degrades the performance. Specifically, performance on Acc7, Acc2 and F1 score has seen a great drop. It is obvious that our proposed contrastive learning method is effective and can greatly boost the performance.

Meanwhile, we design two ablation experiments to investigate the contribution of each component in our proposed HyCon. First, we remove the *refinement loss* from IAMCL and IEMCL (see the case 'W/O Refinement Loss') and the performance of the model has a slight drop. The reason may be that without the extra refinement term in IAMCL and IEMCL, the learning has the probability to fall into a sub-optimal solution where the similarity of positive pairs is not optimized. Fortunately, we can generate a certain amount of negative pairs in each mini-batch, so the degradation is not obvious. The refinement term is still significant as it provides a simple and non-parametric way

TABLE 4
Ablation Studies on the CMU-MOSI Dataset

	Acc7	Acc2	F1	MAE	Corr
HyCon (W/O Contrast Learning)	43.2	82.7	82.8	0.769	0.774
HyCon (W/O Refinement Loss)	45.5	84.9	84.9	0.733	0.789
HyCon (W/O Modality Margin α)	44.2	84.6	84.5	0.739	0.785
HyCon (W/O $\boldsymbol{L}_{IAMCL}$)	45.0	83.2	83.2	0.752	0.787
HyCon (W/O $\boldsymbol{L}_{IEMCL}$)	42.8	83.7	83.4	0.768	0.760
HyCon (W/O $\boldsymbol{L}_{SCL}$)	47.0	84.0	84.0	0.748	0.776
HyCon	46.6	85.2	85.1	0.713	0.790
HyCon-v2 (W/O Positive Selection)	45.3	85.5	85.5	0.714	0.793
HyCon-v2 (W/O Negative Selection)	45.3	**85.7**	**85.6**	0.725	0.794
HyCon-v2	**47.7**	**85.7**	**85.6**	**0.705**	**0.805**

TABLE 5
Discussion on the Fusion Strategies on HyCon

	Acc7	Acc2	F1	MAE	Corr
Concatenation	44.8	84.3	84.1	0.762	0.786
Addition (defaulted)	**46.6**	**85.2**	85.1	**0.713**	0.790
Tensor Fusion [37]	43.7	83.5	83.5	0.763	0.791
Graph Fusion [16]	46.0	85.1	**85.3**	0.729	**0.801**

Graph fusion [16] regards each unimodal, bimodal, and trimodal interaction as one node, and explicitly models their relationship in a parameter-friendly way. Tensor fusion [37] applies outer product to explore interactions, which introduces a large amount of parameters and has high space complexity.

TABLE 6
Discussion on the Selection of Modality Margin α on CMU-MOSI

	Acc7	Acc2	F1	MAE	Corr
HyCon ($\alpha = 0.5$)	42.2	83.8	83.7	0.751	0.784
HyCon ($\alpha = 0.7$)	46.3	83.8	83.8	0.737	**0.795**
HyCon ($\alpha = 0.8$)	**46.6**	**85.2**	**85.1**	**0.713**	0.790
HyCon ($\alpha = 0.9$)	46.1	84.2	84.3	0.747	0.783

to improve the performance of contrastive learning, especially when the negative pairs are rare. Second, we do not consider the *modality margin* α in SCL and IEMCL (see the case 'W/O Modality Margin α' where α is set to 1). The results suggest the necessity to consider α in our loss functions, which may be because modality gap is hard to be completely eliminated, and it is reasonable to allow for retaining the modality-specific information for learning a richer multimodal representation.

Furthermore, we perform ablation study on *three contrastive losses*. It can be seen that removing any contrastive loss degrades the performance of HyCon, which indicates their effectiveness. Specifically, the performance drops significantly when the IAMCL loss or IEMCL loss is removed. It indicates that *the learning of inter-class relationships is the fundamental component that leads to the high performance of our model* (as IAMCL and IEMCL aim to learn intra-/inter-modal inter-class relationships), which should be paid more attention in the research of multimodal learning.

Finally, we perform ablation study on the *pair selection mechanism*. It can be seen that the hard negative mining and hard positive mining are both beneficial to the overall performance of the model, demonstrating the necessity of considering both hard negative and hard positive selection. Interestingly, the hard positive mining is more effective than the hard negative mining. It to some extent suggests that pulling the unimodal representations from the same class closer is more beneficial to the learning of discriminative embedding space, and we should pay more attention to the selection of hard positive samples instead of only focusing on the hard negatives as in previous contrastive learning works[22], [23], [24]. Also notice that we show in Section 4.5.8 that by applying the pair selection mechanism, the running time of the model can be reduced.

4.5.4 Analysis of Fusion Methods

We also conduct experiments to verify that HyCon is generalized to be applied with different fusion strategies. Previous works mostly rely on sophisticated fusion methods to sufficiently learn cross-modal dynamics to reach satisfactory results. Unlike them, our HyCon can achieve state-of-the-art performance with simple fusion strategies. As shown in Table 5, *even with simple and direct fusion methods like concatenation and element-wise addition of unimodal representations, HyCon still reaches competitive performance*. Note that *all the four HyCon variants achieve favorable performance compared to baselines*. A conclusion can be reached that our designed hybrid contrastive learning is effective and of satisfactory generalization ability.

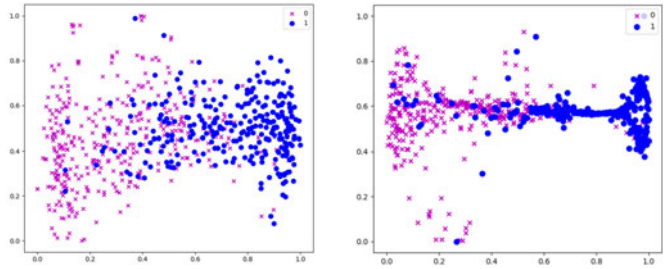
Moreover, we can notice that the direct addition fusion performs best, while the tensor fusion performs worst. We argue that it is because by using the cross-modal contrastive learning, the modality gap can be significantly reduced. And the feature dimensionality of unimodal and multimodal representations are forced to have approximately the same distribution, such that direct addition is strong enough to explore the complementary information and interactions between modalities. Instead, by applying outer product to explore interactions, the tensor fusion [37] introduces a large amount of parameters and has high complexity, which may introduce noise to the feature distribution and degrade the performance of the multimodal system. Moreover, as a parameter-efficient model, graph fusion [16] regards each unimodal, bimodal, and trimodal interaction as one node, and explicitly models their relationship, which also achieves high performance.

4.5.5 Discussion on the Selection of Modality Margin α

Having verified the effectiveness of considering the modality margin α, we further carry out experiments to investigate the effect of different selections of its value. Our HyCon achieves the best performance when α is set to be 0.8. From Table 6, we can see that our model suffers degradation when α is set to be either higher or lower than 0.8. If α is set to be too low (in the case when $\alpha = 0.5/0.7$), there exists large modality gap which hinders higher performance. On the contrary, if α is too high (in the case when $\alpha = 0.9$), modality-specific information may be lost to align different modalities. All in all, the selection of α is of great significance to ensure an optimal solution.

4.5.6 Visualization for the Embedding Space

We provide a visualization for distributions of mutlimodal representation in the embedding space where the right and left sub-figure on Fig. 4 illustrate the embedding space learnt by HyCon and learnt without contrastive learning, respectively. The visualization result of HyCon-v2 is similar to that of HyCon, and is omitted here. The visualization is obtained by transforming the multimodal representation into 2-dimensional feature point using t-SNE algorithm. We can infer from Fig. 4 that when the contrastive losses are removed, the data points in the embedding space tend to be very scattered, and different classes does not form a distinguishing cluster. Instead, when the contrastive losses are added to the model, the distance between the data points is significantly narrowed down, and each sentiment class form a discriminative cluster. Moreover, the centers of the



(a) Learned without contrastive learning (b) Learned by HyCon

Fig. 4. *T-SNE visualization of the embedding space.* The 'purple x' and 'blue dot' denote the data point for negative and positive sentiments, respectively.

two sentiment clusters are far away, and the data points that are difficult to be clustered are in the middle of the embedding space. This is because we explicitly model the distance (similarity) of different samples to pull the samples of the same class to be closer and push the samples of different classes to be further apart, which is helpful for the classifier to make prediction. Nevertheless, there are some dots that are extremely difficult to be correctly classified found in the wrong clusters. This is reasonable because the accuracy of the sentiment classification is only about 85% even with a well-trained classifier.

4.5.7 Comparison of Loss Functions

In this section, we compare the proposed contrastive losses with other loss functions to analyze the effectiveness. The candidate loss functions include the widely-used triplet loss [59], N-pair loss [60], and the classical contrastive loss that only considers one positive pair for each anchor. The equations for these losses are presented as follows:

$$\boldsymbol{L}_{con} = -E_s\left[\frac{\boldsymbol{a}^T\boldsymbol{p}}{\boldsymbol{a}^T\boldsymbol{p} + \sum_{j=1}^{M}\boldsymbol{a}^T\boldsymbol{n}_j}\right] \tag{18}$$

$$\boldsymbol{L}_{tri} = E_s\left[||\boldsymbol{a}-\boldsymbol{p}||_2^2 - ||\boldsymbol{a}-\boldsymbol{n}||_2^2 + 1\right] \tag{19}$$

$$\boldsymbol{L}_{n-pair} = E_s\left[log(1+\sum_{j=1}^{m} exp(\boldsymbol{a}^T\boldsymbol{n}_j - \boldsymbol{a}^T\boldsymbol{p}))\right] \tag{20}$$

where $\boldsymbol{a}$ is the anchor, $\boldsymbol{n}$ is the negative sample of the anchor, and $\boldsymbol{p}$ is the positive sample of the anchor. $\boldsymbol{L}_{con}$, $\boldsymbol{L}_{tri}$, and $\boldsymbol{L}_{n-pair}$ denotes the classical contrastive loss, triplet loss, and N-pair loss, respectively.

We use the above-mentioned losses to replace the proposed $\boldsymbol{L}_{IAMCL}$ and $\boldsymbol{L}_{IEMCL}$, and the settings of the negative and positive pairs remain the same. However, triplet loss only considers one positive pair and one negative pair where the pairs are randomly sampled from the mini-batch. N-pair loss and the classical contrastive loss consider one positive pair and many negative pairs where the positive pair is randomly sample from the mini-batch and the negative pairs are the same as those of $\boldsymbol{L}_{IAMCL}$ and $\boldsymbol{L}_{IEMCL}$. One advantage of the supervised contrastive learning is that it can leverage many positive and negative samples for each anchor, which avoids the need of hard-negative mining. We also compare the proposed losses with hard-triplet mining, which has the same equation with the classical triplet loss, but considers hard positive and hard negative pairs. The chosen hard positive sample shares the lowest similarity of anchor among all positive samples in the mini-batch and the hard negative sample shares the highest similarity of anchor.

TABLE 7
Comparison of Loss Functions

	Acc7	Acc2	F1	MAE	Corr
Triplet Loss	43.4	83.5	83.4	0.797	0.777
Hard-triplet Mining	45.5	84.3	84.2	0.753	0.776
N-pair Loss	45.7	83.8	83.9	0.767	0.768
Classical Contrastive Loss	46.3	83.7	83.6	0.740	0.791
HyCon	46.6	85.2	85.1	0.713	0.790
HyCon-v2	**47.7**	**85.7**	**85.6**	**0.705**	**0.805**

From Table 7, we can infer that the hard-triplet mining loss performs better than the classical triplet loss significantly on Acc7, Acc2, F1 score and MAE, and reach a similar performance on Corr metric. These results suggest the effectiveness of hard negative/positive mining for triplet loss. However, our HyCon still outperforms hard-triplet mining by a significant margin on all the metrics, mainly for the reason that HyCon considers many positive and many negative pairs for training which is more favorable than one hard negative/positive pair. As for N-pair loss and classical contrastive loss, they both reach satisfactory performance and outperform the classical triplet loss, possibly for the reason that they consider multiple negative pairs. Nevertheless, HyCon performs better than them by a large margin for it generates multiple positive pairs for each anchor, and HyCon-v2 outperforms HyCon for the reason than it considers hard negative and hard positive mining. Moreover, HyCon considers modality margin to learn cross-modal mapping, which allows the existence of modality-specific information of each modality for fusion and thereby achieves more favorable performance.

Note that by applying any one of these loss functions, our method can outperform the model without contrastive losses (see Table 4 for the results of model without contrastive learning), further demonstrating that the idea of exploring cross-modal and modality-specific dynamics between intra- and inter-class samples is effective.

4.5.8 Comparison of Running Time

In this section, we compare the model complexity between the HyCon and the baselines. To estimate the space complexity, we use the number of parameters as the proxy. For the evaluation of time complexity, we use the running time as the proxy (all the methods use the same batch size and the same number of epochs). As shown in Table 8, the running time of our model is slightly longer than that of the majority of baselines, for the reason that our HyCon generates a large number of positive and negative pairs for training. HyCon-v2 runs slightly faster than the original HyCon, indicating that the pair selection mechanism helps to reduce the computational overhead. The MFN-BERT has the longest running time for the reason that it applies lots of recurrent neural networks which cannot operate in parallel in the time dimension. Also note that HyCon's inference time

TABLE 8
The Comparison of Model Complexity

	The number of parameters	Training Time (s)
MAG-BERT [12]	118,679,809	293
TFN-BERT [37]	161,409,399	275
GFN-BERT [16]	109,232,350	274
MFN-BERT [4]	111,256,325	455
MULT-BERT [10]	110,838,399	378
HyCon	109,209,149	440
HyCon-v2	109,209,149	429

during testing is much faster as the negative and positive pairs are not generated during testing and the fusion method is very simple. It takes HyCon-v2 about 0.770s to finish testing 686 utterances on the CMU-MOSI testing set. Therefore, our method is applicable to large datasets. For the number of parameters, our HyCon has fewer parameters than the baselines, for the reason that our multimodal contrastive learning framework does not introduce additional parameters.

5 Conclusion

We propose a novel framework HyCon for hybrid contrastive learning of cross-modal representations to perform multimodal sentiment analysis, which is capable of learning intra-/inter-modal dynamics while at the same time exploring inter-class relationships. Specifically, our proposed HyCon consists of three ways of contrastive learning, i.e., intra-modal contrastive learning, inter-modal contrastive learning and semi-contrastive learning. In this way, the designed loss function with a refinement term allows the model to sufficiently learn latent embeddings for sentiment prediction in an optimal way. Moreover, learning with our designed loss function introduce no additional parameters so as to reduce the possibility of overfitting and improve the generalization ability. Experiments demonstrate that our HyCon outperforms state-of-the-art methods. Notably, our work focuses on the contrastive learning between unimodal representations, but pays little attention to the design of novel fusion methods. In the future, we aim to inject the popular graph-based learning strategies [16], [41], [61] into our framework. Specifically, using the graph construction methods [62], [63], [64], [65] to construct the multimodal graph is beneficial for learning the connections and interactions between modalities, improving the explanation ability of multimodal fusion. Moreover, the weight of the selected pairs in HyCon-v2 can be improved by exploring the idea in [66].

Acknowledgments

Sijie Mai and Ying Zeng are contributed equally to this work.

References

[1] 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.
[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[3] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2014, pp. 1724–1734.
[4] 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.
[5] S. Mai, H. Hu, and S. Xing, "Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 481–492.
[6] 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.
[7] S. Mai, S. Xing, and H. Hu, "Locally confined modality fusion network with a global perspective for multimodal human affective computing," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 122–137, Jan. 2020.
[8] S. Mai, H. Hu, J. Xu, and S. Xing, "Multi-fusion residual memory network for multimodal human sentiment comprehension," *IEEE Trans. Affective Comput.*, vol. 13, no. 1, pp. 320–334, Jan.–Mar. 2022.
[9] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[10] 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. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6558–6569.
[11] Y.-H. H. Tsai, M. Q. Ma, M. Yang, R. Salakhutdinov, and L.-P. Morency, "Multimodal routing: Improving local and global interpretability of multimodal language analysis," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, Art. no. 1823.
[12] W. Rahman *et al.*, "Integrating multimodal information in large pretrained transformers," in *Proc. Conf. Assoc. Comput. Linguistics Meeting*, 2020, pp. 2359–2369.
[13] 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.
[14] K. Yang, H. Xu, and K. Gao, "CM-BERT: Cross-modal bert for text-audio sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 521–528.
[15] A. Zadeh, R. Zellers, E. Pincus, and L. P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov. 2016.
[16] S. Mai, H. Hu, and S. Xing, "Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 164–172.
[17] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[18] K. Hassani and A. H. Khasahmadi, "Contrastive multi-view representation learning on graphs," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 4116–4126.
[19] P. Khosla *et al.*, "Supervised contrastive learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 18 661–18 673.
[20] V. Udandarao, A. Maiti, D. Srivatsav, S. R. Vyalla, Y. Yin, and R. Shah, "COBRA: Contrastive Bi-modal representation algorithm," 2020, *arXiv:2005.03687*.
[21] J. D. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka, "Contrastive learning with hard negative samples," in *Proc. Int. Conf. Learn. Representations*, 2021.
[22] W. Zhang and K. Stratos, "Understanding hard negatives in noise contrastive estimation," 2021, *arXiv:2104.06245*.
[23] T. T. Cai, J. Frankle, D. J. Schwab, and A. S. Morcos, "Are all negatives created equal in contrastive instance discrimination?" 2020, *arXiv:2010.06682*.
[24] M. Wu, M. Mosse, C. Zhuang, D. Yamins, and N. Goodman, "Conditional negative sampling for contrastive learning of visual representations," in *Proc. Int. Conf. Learn. Representations*, 2021.
[25] 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. Affective Comput.*, to be published, doi: 10.1109/TAFFC.2021.3097002.
[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. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253520303365

[27] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, 2017.
[28] J. Wagner, E. Andre, F. Lingenfelser, and J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data," *IEEE Trans. Affective Comput.*, vol. 2, no. 4, pp. 206–218, Oct.–Dec 2011.
[29] M. Wöllmer *et al.*, "YouTube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 46–53, May/Jun. 2013.
[30] V. Rozgic, S. Ananthakrishnan, S. Saleem, and R. Kumar, "Ensemble of SVM trees for multimodal emotion recognition," in *Proc. Signal Inf. Process. Assoc. Summit Conf.*, 2012, pp. 1–4.
[31] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proc. IEEE Int. Conf. Data Mining*, 2016, pp. 439–448.
[32] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L. P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 873–883.
[33] C.-H. Wu and W.-B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Trans. Affective Comput.*, vol. 2, no. 1, pp. 10–21, Jan.–Jun. 2011.
[34] B. Nojavanasghari, D. Gopinath, J. Koushik, and L. P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2016, pp. 284–288.
[35] O. Kampman, E. J. Barezi, D. Bertero, and P. Fung, "Investigating audio, visual, and text fusion methods for end-to-end automatic personality prediction," in *ACL Short Paper*, 2018, *arxiv:1805.00705*.
[36] Y. Zhao, X. Cao, J. Lin, D. Yu, and X. Cao, "Multimodal affective states recognition based on multiscale CNNs and biologically inspired decision fusion model," *IEEE Trans. Affective Comput.*, to be published, doi: 10.1109/TAFFC.2021.3093923.
[37] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2017, pp. 1114–1125.
[38] M. Hou, J. Tang, J. Zhang, W. Kong, and Q. Zhao, "Deep multimodal multilinear fusion with high-order polynomial pooling," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 12 113–12 122.
[39] Z. Liu, Y. Shen, P. P. Liang, A. Zadeh, and L. P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.
[40] H. Pham, P. P. Liang, T. Manzini, L. P. Morency, and P. Barnabăs, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6892–6899.
[41] A. Zadeh *et al.*, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.
[42] 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.
[43] T. Pan, Y. Song, T. Yang, W. Jiang, and W. Liu, "VideoMoCo: Contrastive video representation learning with temporally adversarial examples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11 200–11 209.
[44] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 776–794.
[45] S. Ma, Z. Zeng, D. McDuff, and Y. Song, "Active contrastive learning of audio-visual video representations," in *Proc. Int. Conf. Learn. Representations*, 2021.
[46] D. Olson, "From utterance to text: The bias of language in speech and writing," *Harvard Educ. Rev.*, vol. 47, no. 3, pp. 257–281, 1977.
[47] 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: Human Lang. Technol.*, 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19–1423
[48] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, and D. Larlus, "Hard negative mixing for contrastive learning," 2020, *arXiv:2010.01028*.
[49] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.
[50] S. Mai, S. Xing, and H. Hu, "Analyzing multimodal sentiment via acoustic- and visual-LSTM with channel-aware temporal convolution network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1424–1437, 2021.
[51] 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.
[52] 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.
[53] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, Art. no. 517.
[54] D. Gkoumas, Q. Li, C. Lioma, Y. Yu, and D. W. Song, "What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis," *Inf. Fusion*, vol. 66, pp. 184–197, 2021.
[55] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14–1162
[56] 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.
[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.
[58] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov./Dec. 2016.
[59] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, 2009.
[60] K. Sohn, "Improved deep metric learning with multi-class N-pair loss objective," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 1857–1865.
[61] S. Mai, S. Xing, J. He, Y. Zeng, and H. Hu, "Analyzing unaligned multimodal sequence via graph convolution and graph pooling fusion," 2020, *arXiv:2011.13572*.
[62] Z. Li, F. Nie, X. Chang, Y. Yang, C. Zhang, and N. Sebe, "Dynamic affinity graph construction for spectral clustering using multiple features," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6323–6332, Dec. 2018.
[63] M. Angelou, V. Solachidis, N. Vretos, and P. Daras, "Graph-based multimodal fusion with metric learning for multimodal classification," *Pattern Recognit.*, vol. 95, pp. 296–307, 2019.
[64] Z. Li, F. Nie, X. Chang, L. Nie, H. Zhang, and Y. Yang, "Rank-constrained spectral clustering with flexible embedding," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6073–6082, Dec. 2018.
[65] M. Behmanesh, P. Adibi, M. S. Ehsani, and J. Chanussot, "Geometric multimodal deep learning with multi-scaled graph wavelet convolutional network," 2021, *arXiv:2111.13361*.
[66] Z. Li, L. Yao, X. Chang, K. Zhan, J. Sun, and H. Zhang, "Zero-shot event detection via event-adaptive concept relevance mining," *Pattern Recognit.*, vol. 88, pp. 595–603, 2019.