Dual-View Learning for Conversational Emotion Recognition Through Context and Emotion-Shift Modeling

Xupeng Zha, Huan Zhao*, Guanghui Ye, Zixing Zhang*

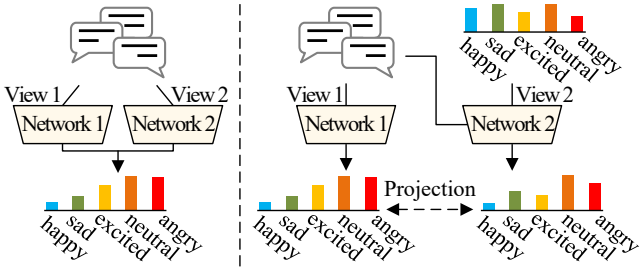
College of Computer Science and Electronic Engineering, Hunan University, China
{zhaxupeng, hzhao, yghui, zixingzhang}@hnu.edu.cn

Abstract

Conversational Emotion Recognition (CER) has recently been explored through conversational context modeling to learn the emotion distribution, i.e., the likelihood over emotion categories associated with each utterance. While these methods have shown promising results in emotion classification, they often focus on the interactions between utterances (utterance-view) and overlook shifts in the speaker's emotions (emotion-view). This emphasis on homogeneous view modeling limits their overall effectiveness. To address this limitation, we propose DVL-CER, a novel **D**ual-**V**iew **L**earning approach for **CER**. DVL-CER integrates both the utterance-view and emotion-view using two projection heads, enabling cross-view projection of emotion distributions. Our approach offers several key advantages: (1) We introduce an emotion-view that captures shifts in a speaker's emotions from initial to subsequent states within a conversation. This view enriches the conversation modeling and supports seamless integration with various CER baseline models. (2) Our dual-view projection learning strategy flexibly balances consistency and independence between the two heterogeneous views, promoting view-specific adaptation learning and incorporating the emotion verification capability within CER. We validate DVL-CER through extensive experiments on two widely-used datasets, IEMOCAP and EmoryNLP. The results demonstrate that DVL-CER achieves state-of-the-art performance, delivering robust and high-quality emotion distributions compared with existing CER methods and other dual-view learning strategies.

1 Introduction

Conversational Emotion Recognition (CER) is a crucial task in natural language processing that involves identifying the emotion expressed in each utterance within a dialogue. With the increasing prevalence of conversational data and the growing demand for empathetic AI systems, CER has gained significant importance. This research is essential for applications, such as social media opinion analysis (Kumar, Dogra, and Dabas 2015) and emotion-aware chatbots (Chatterjee et al. 2019), highlighting its importance in understanding and recognizing emotional states in conversations.



(a) Existing dual-view learning methods. (b) Our dual-view projection learning method.

Figure 1: Comparison between our dual-view projection learning method for CER and existing dual-view learning methods (e.g., Ma et al. (2022) and Ruan et al. (2022)) in terms of view choice and view sharing.

Current CER research focuses heavily on predicting the emotion of each utterance by considering both intra- and inter-speaker emotion dependencies within the conversational context. Common methods employ Recurrent Neural Networks (RNNs) (Majumder et al. 2019; Ghosal et al. 2020; Hu, Wei, and Huai 2021) or Graph Neural Networks (GNNs) (Ghosal et al. 2019; Li et al. 2021; Shen et al. 2021b) to model these dependencies, transforming the semantic content of utterances into emotion representations for learning emotion distributions, i.e., the likelihood over emotion categories for each utterance. Despite some successes, these methods face challenges, including interaction discord and transformation inconsistency. These issues stem primarily from two factors: i) natural language semantics in utterance embeddings do not adequately encode affective meaning due to the distributional hypothesis (Faruqui et al. 2015; Babanejad et al. 2024); and ii) the learned emotion distributions, though effective for emotion prediction, may lack sufficient sensitivity to capture subtle emotional nuances (Felbo et al. 2017; Poria et al. 2020). Consequently, the current focus on context modeling from an utterance perspective limits the potential to capture emotion dependencies.

To overcome these challenges, previous efforts have introduced external knowledge to enhance emotion comprehension in models. For instance, methods like KET (Zhong, Wang, and Miao 2019), which incorporates the NRC VAD emotion lexicon (Mohammad 2018), and COSMIC (Ghosal et al. 2020), which fine-tunes a pre-trained RoBERTa

*Corresponding authors.

model (Liu et al. 2019) for emotion label prediction, aim to embed affective meaning into utterance embeddings. On the other hand, approaches like SKAIG (Li et al. 2021), COSMIC, and CauAIN (Zhao, Zhao, and Lu 2022) leverage commonsense knowledge from COMET (Bosselut et al. 2019) to mediate interactions between utterances, thereby improving information transfer and the model's expressive capabilities. However, despite these advances, existing studies have not fully considered and explored emotion dependencies from an emotion-centric perspective. Thus, it is necessary to investigate emotion view modeling to capture these dependencies and interactions more effectively.

Emotion shift, a concept extensively studied in the CER literature, is defined in psychology as the "evolution of emotional experience during exposure to a media message" (Nabi and Green 2015). Current research (Ghosal et al. 2021; Yang et al. 2022; Gao et al. 2022; Bansal et al. 2022; Li et al. 2024) focuses primarily on emotion label shifts rather than on the underlying emotion dynamics. Yang et al. (2022) and Gao et al. (2022) investigate the consistency of emotion representation across consecutive utterances by the same speaker, while other studies (Bansal et al. 2022; Li et al. 2024) explore the probability of emotion label shifts between consecutive or any two utterances, regardless of speaker identity. However, these studies typically address only emotion variables and binary classification problems, neglecting the media message—the sequence of utterances between two consecutive emotions from the same speaker, which limits understanding of emotion shifts. To address this gap, we propose an emotion shift modeling approach that integrates the initial emotion with the utterance sequence to predict future emotions, aiming to explicitly reveal intra-speaker emotion transfer and interactions.

Building on both the conversational context network (utterance-view) and the emotion shift network (emotion-view), we propose a hypothesis that the emotion distributions learned from these two views are shared in CER. Under this hypothesis of emotion distribution consistency, we introduce a dual-view projection learning strategy, where two projection heads predict the emotion distributions of one view from another view. This strategy ensures consistency between views by aligning cross-view emotion distributions while preserving view-specific independence. We call the resulting framework DVL-CER, a **D**ual-**V**iew **L**earning approach for **CER**, which, to the best of our knowledge, is the first to model an emotion view and to explore category distribution consistency under dual/multi-view supervision learning. This approach distinguishes our work from existing CER solutions, which rely solely on homogeneous utterance-view modeling and representation-level view sharing, as illustrated in Figure 1. Additionally, our framework is agnostic to the conversational context network and can be seamlessly integrated with it.

Our main contributions can be summarized as follows:

- We introduce DVL-CER, a novel dual-view learning framework for CER that enables desirable sharing of emotion distributions between a conversational context network for utterance-view modeling and an emotion shift network for emotion-view modeling through two projection heads.
- DVL-CER exhibits strong generalization ability, allowing seamless integration with various CER baselines.
- Our dual-view projection learning strategy effectively balances the trade-off between consistency and independence across two heterogeneous views.
- Extensive experiments on the IEMOCAP and EmoryNLP benchmarks show that DVL-CER produces more robust and higher-quality emotion distributions compared with existing CER methods and other dual-view learning strategies, achieving state-of-the-art results.

2 Related Works

From the perspective of model architecture, CER methods can be broadly categorized into two groups: single-view and multi-view CER. Multi-view CER methods employ multiple networks, each tailored to a specific interaction perspective, allowing for fine-grained modeling of conversational context. In contrast, single-view CER methods utilize a single network to model the interaction relationships considered within a conversation, simplifying the learning setting.

Single-View CER. Single-view CER methods often employ GNN-based frameworks, representing the conversation as a graph with utterances as nodes, speaker dependencies as edges, and dependency types as edge attributes. Notable examples include DialogueGCN (Ghosal et al. 2019), which models both intra- and inter-speaker dependencies; SKAIG (Li et al. 2021), which enhances edge representations with psychological knowledge; and DialogXL (Shen et al. 2021a), which uses a self-attention mechanism to capture crucial dependencies between speakers.

Multi-View CER. Multi-view CER methods leverage multiple networks to model different interaction perspectives of conversational context. While commonly applied in multimodal CER (Zadeh et al. 2018; Meng et al. 2024), where each modality corresponds to a distinct perspective, multi-view learning is also present in textual CER. For example, HMVDM (Ruan et al. 2022) introduces a hierarchical structure to capture token-level and utterance-level dependencies, and MVN (Ma et al. 2022) proposes a dual-view learning model that combines an attention mechanism for word-level dependencies with a bidirectional gated recurrent unit for utterance-level dependencies. These methods, despite being termed as "multi-view learning," share similarities with CER approaches that employ multiple networks for modeling different dependencies or perspectives. Examples include DialogueRNN (Majumder et al. 2019), which models speaker, context, and emotion states with three GRUs; COSMIC (Ghosal et al. 2020), which refines speaker states—context, internal, external, intent, and emotion—using five GRUs; DialogueCRN (Hu, Wei, and Huai 2021), which captures situation-level and speaker-level context with two LSTMs; and DualRAN (Li, Wang, and Zeng 2024), which employs a dual-stream recurrence-attention network to extract contextual information. These methods are, therefore, categorized as multi-view learning methods.

As alluded to earlier, both single-view and multi-view CER methods emphasize conversational context modeling

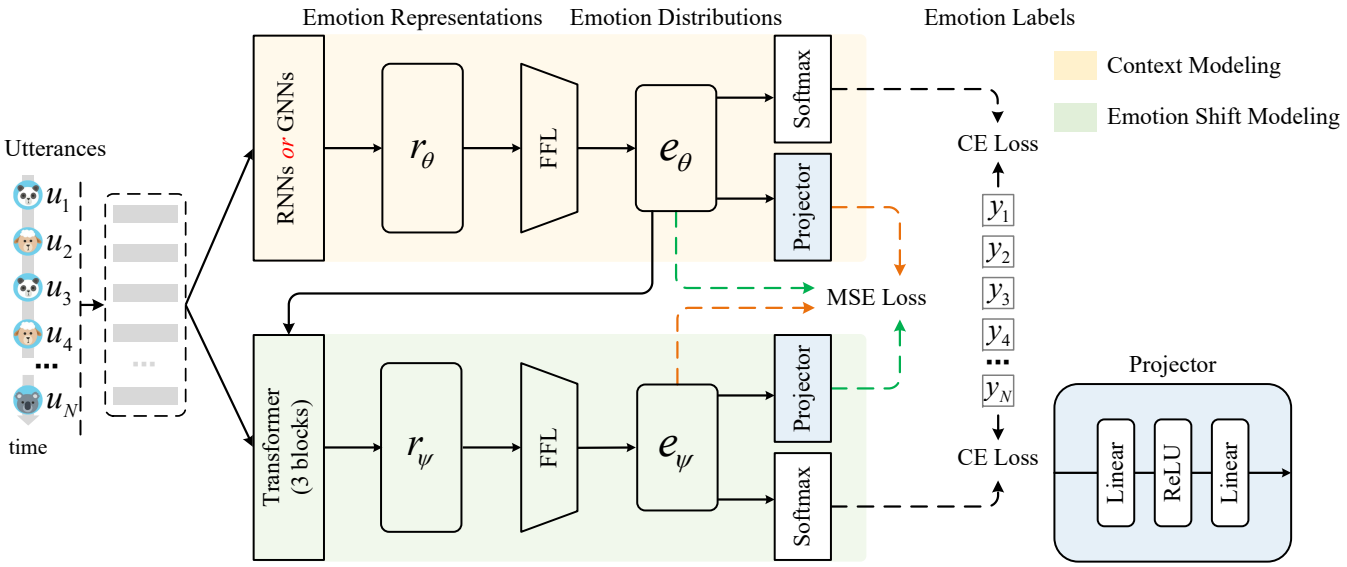


Figure 2: Illustration of the proposed DLV-CER approach. This approach employs a base network—either based on recurrence or graph networks—to model conversational context (utterance-view) and a Transformer-based network to capture emotion shifts (emotion-view), with both networks aimed at learning emotion distributions. Two projection heads then map these emotion distributions from one view to another, optimized using the Mean Squared Error (MSE) loss function. At the end of training, the emotion-view network and two projection heads are discarded, leaving the utterance-view network and its learned emotion distribution e_θ for emotion classification. MSE loss functions, indicated by dotted lines of the same colors, are computed.

from an utterance perspective. In contrast, this paper introduces the first emotion-view network, focusing on modeling emotion shifts and integrating existing CER networks into a novel heterogeneous dual-view CER framework.

Multi-View Learning Strategies. In recent decades, multi-view learning (Sun 2013; Yan et al. 2021) has gained considerable attention in machine learning and computer vision, inspiring many promising algorithms. Prominent among these are multi-view alignment learning (Radford et al. 2021), multi-view feature aggregation learning (Yang et al. 2018), and multi-view subspace learning (Andrew et al. 2013; Xue et al. 2019). The latter approach, which searches for two projections to filter out view-specific independence while mapping views onto a common low-dimensional subspace to maximize correlation, is particularly relevant to our work. In contrast to this filtering mechanism, our proposed multi-view projection learning strategy aims to find two projections that map one view onto the space of another, maximizing correlation while leaving view-specific independence. Furthermore, rather than conventional representation-level sharing, our method strives to a direct sharing that targets task-relevant category distributions.

3 Methodology

3.1 Task Definition

We define the task of CER as follows: Given a conversation consisting of N consecutive utterances $U = [u_1, u_2, \ldots, u_N]$ spoken by M speakers $S = [s_1, s_2, \ldots, s_M]$, the objective is to learn the emotion distribution e_i for each utterance u_i, with the corresponding emotion label $y_i \in \{y_1, y_2, \ldots, y_C\}$ serving as supervision. In line with prior work by Ghosal et al. (2020), we use the RoBERTa Large model (Liu et al. 2019) to extract an utterance embedding $h_i \in \mathbb{R}^d$ for each utterance u_i.

The proposed DVL-CER framework comprises three main components: an utterance-view network for modeling conversational context, an emotion-view network for modeling emotion shifts, and two projectors that facilitate cross-view correlations, as illustrated in Figure 2.

3.2 Conversational Context Modeling

Traditional CER methods often model conversational context from an utterance view to address the task defined earlier. These methods typically employ RNNs or GNNs to transform the utterance embeddings $H \in \mathbb{R}^{N \times d}$ into emotion representations r_θ. These representations are subsequently processed by a Feed-Forward Layer (FFL) to predict emotion distributions e_θ, guided by the corresponding emotion labels. The process is formalized as:

$$e_\theta = \text{FFL}(\text{X-NNs}(H)), \tag{1}$$

where $e_{\theta,i} \in \mathbb{R}^{|C|}$ represents the predicted emotion distribution for utterance u_i, and X-NNs refers to either RNNs or GNNs. The conversational context network is trained using a Cross-Entropy (CE) loss function:

$$\mathcal{L}_{CCM} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{|C|} y_i^c \log(\text{softmax}(e_{\theta,i})), \tag{2}$$

where y_i^c is a one-hot vector for the emo-tion label of utterance u_i, and c denotes the dimension of each emotion label.

Since the subsequent sections focus on the final emotion

distribution e_θ without directly involving the conversational context network, our framework accommodates various choices for the conversational context network without imposing any constraints.

3.3 Emotion Shift Modeling

While emotion distributions derived from conversational context modeling are often sufficient for emotion classification, they may lack explicit interactions between emotions, leading to potential inconsistencies and biases that affect prediction accuracy. To address this, we propose emotion shift modeling, which explicitly captures these interactions.

Given a conversation segment $U_w = [u_i, \ldots, u_{i+w-1}] \in U$ and an emotion distribution $e_{\theta,i}$ from the conversational context network, where u_i and u_{i+w-1} are spoken by the same speaker, i.e., $s_i = s_{i+w-1}$, and the intermediate utterances $[u_{i+1}, \ldots, u_{i+w-2}]$ are from others, with w representing the emotion shift span (i.e., the number of utterances between the speaker's initial and future emotion), we treat $e_{\theta,i}$ as the initial emotion and U_w as the medium through which the emotion shift occurs. A Transformer-based model is used to predict the future emotion label y_{i+w-1}.

First, the emotion distribution $e_{\theta,i}$ is embedded into a vector $h^e_{\theta,i}$ of dimension $\mathbb{R}^d$ through two fully connected layers:

$$h^e_{\theta,i} = W_2(W_1 e_{\theta,i} + b_1) + b_2. \tag{3}$$

Here, W_1, W_2, b_1, and b_2 are trainable parameters.

Next, a special token, [CLS], is prepended to the sequence $[h^e_{\theta,i}, h_i, \ldots, h_{i+w-1}]$ to serve as an aggregator for both the initial emotion and the medium, forming an input sequence $z_i = [h_{[\text{CLS}]}, h^e_{\theta,i}, h_i, \ldots, h_{i+w-1}] \in \mathbb{R}^{(w+2)\times d}$. To account for the temporal nature of the emotion shift, position embeddings $p_t \in \mathbb{R}^{(w+2)\times d}$, encoded by sine and cosine functions, are added to the input sequence:

$$z^0_i = z_i + p_t. \tag{4}$$

The input sequence $z^0_i \in \mathbb{R}^{(w+2)\times d}$ is then processed through a standard Transformer architecture with L layers. The l-th Transformer block is defined as follows:

$$\tilde{g}^l_i = \textsc{MultiAttn}(z^{l-1}_i), \tag{5}$$

$$g^l_i = \textsc{LayerNorm}(\tilde{g}^l_i + z^{l-1}_i), \tag{6}$$

$$\tilde{z}^l_i = \text{FFN}(g^l_i), \tag{7}$$

$$z^l_i = \textsc{LayerNorm}(\tilde{z}^l_i + g^l_i), \tag{8}$$

where MULTIATTN represents the multi-head self-attention mechanism, LAYERNORM refers to layer normalization, and FFN is a two-layer feed-forward neural network with ReLU activation. After passing through L layers, the hidden state $z^L_{i,[\text{CLS}]} \in \mathbb{R}^d$ of the [CLS] token from the final layer is used as the future emotion representation $r_{\psi,i+w-1}$.

Finally, these emotion representations r_ψ are encoded into an emotion distributions e_ψ through a Feed-Forward Layer:

$$e_\psi = \text{FFL}(r_\psi). \tag{9}$$

The loss function for emotion shift modeling is as follows:

$$\mathcal{L}_{ESM} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{e=1}^{|C|} y^e_i \log(\text{softmax}(e_{\psi,i})). \tag{10}$$

In our approach, we use the emotion distribution $e_{\theta,i}$ as the initial emotion instead of the emotion label y_i, forming an implicit emotion validation chain, which we will discuss later. Empirical evidence indicates that this chain enhances model performance. Moreover, leveraging labeled data in single-view learning helps guide representations and distributions toward task-relevant features (Zhai et al. 2019; Henaff 2020).

3.4 Dual-View Projection Learning

Our objective is to develop robust emotion distributions that can effectively classify emotions across conversations while capturing shifts within individual emotions. This task is challenging due to the significant discrepancy between the heterogeneous views involved. A straightforward solution is to use existing multi-view learning strategies that aggregate information from multiple views by maximizing correlation or independence. However, these strategies, as demonstrated in Section 4.4, are effective but generally optimized for representation rather than category distribution, leading to suboptimal results. To address this limitation, we propose a dual-view projection learning strategy tailored to emotion distributions in CER. Inspired by contrastive learning (Chen et al. 2020; Grill et al. 2020), our approach predicts different views of the same utterance from one another, enabling the sharing of learned emotion distributions between views.

Given two emotion distributions, e_θ and e_ψ, derived from the utterance and emotion views, respectively, we recognize that these distributions exist in distinct, heterogeneous spaces. To bridge this gap, we employ two non-linear projectors, f_θ and f_ψ, to directly learn the output of one view from the other. Consistency between views and projections is then calculated within their respective spaces. The view-sharing objective is defined using the Mean Squared Error (MSE) loss function:

$$\mathcal{L}^\theta_{MSE} = ||e_\theta - f_\psi(e_\psi)||^2_2, \tag{11}$$

$$\mathcal{L}^\psi_{MSE} = ||e_\psi - f_\theta(e_\theta)||^2_2. \tag{12}$$

These projectors provide information from other views and translate the individually training process within each view's space into a process of mutual learning and interaction. This approach maximizes cross-view correlations without sacrificing the independence of each view.

Emotion Validation Chain. Our framework further benefits from the emotion shift network, which inputs e_θ and utilizes the projector f_ψ to output e_θ. This creates an implicit emotion validation chain, forming a feedback loop that injects the model's ability to validate e_θ.

Why Emotion Distribution Instead of Emotion Representation? The distinction lies in information density. Emotion representations carry more detailed information than emotion distributions, making cross-view prediction more challenging. Additionally, the continuous updating of initial emotions can lead to instability and complicate future emotion predictions. Therefore, our approach formulates tasks involving emotion distributions: (1) The FFL compresses high-dimensional emotion representations into task-specific low-dimensional emotion distributions, akin to knowledge

Datasets	#Dialogues			#Utterances			#Classes
	train	val	test	train	val	test	
IEMOCAP	120		31	4,810	1,000	1,623	6
EmoryNLP	713	99	85	9,934	1,344	1,328	7

Table 1: Statistics of the datasets.

distillation (Gou et al. 2021), where logits act as carriers of view information. (2) Small deviations in emotion distribution vectors (e.g., six classes in IEMOCAP) do not result in significant projection losses, contributing to training stability. (3) Unlike unsupervised and semi-supervised representation learning, which attempts to learn a generalizable representation, our fully supervised approach directly learns a class distribution tailored to the classification task.

3.5 Training and Inference

Training. We train the DVL-CER approach using a combination of cross-entropy losses $\mathcal{L}_{CCM}$ and $\mathcal{L}_{ESM}$ for single-view emotion classification, and MSE losses $\mathcal{L}^{\theta}_{MSE}$ and $\mathcal{L}^{\psi}_{MSE}$ for cross-view correlation:

$$\mathcal{L} = \mathcal{L}_{CCM} + \mathcal{L}_{ESM} + \lambda_1 \mathcal{L}^{\theta}_{MSE} + \lambda_2 \mathcal{L}^{\psi}_{MSE}, \quad (13)$$

where λ_1 and λ_2 are hyperparameters that balance the trade-off between independence ($\mathcal{L}_{CCM}$ and $\mathcal{L}_{ESM}$) and correlation ($\mathcal{L}^{\theta}_{MSE}$ and $\mathcal{L}^{\psi}_{MSE}$).

Inference. The resulting emotion distributions from both views are expected to aggregate all information and achieve high distribution consistency (see Section 4.6), despite residing in different feature spaces. For inference, we retain only the conversational context network and its corresponding emotion distribution e_{θ}, ensuring comparability with existing CER baselines and addressing two key questions:

Question I: Is the emotion shift modeling scheme reasonable and effective?

Question II: Does the proposed dual-view projection learning strategy effectively integrate information from other views, enhancing single-view learning and its emotion distributions?

4 Experiments

4.1 Experiment Setup

Datasets and Evaluation Metric. We evaluate our DVL-CER approach using two benchmark datasets: IEMOCAP (Busso et al. 2008) and EmoryNLP (Zahiri and Choi 2018). Detailed descriptions of these datasets are provided in Appendix A[1], and the statistics are presented in Table 1.

Following recent studies (Ghosal et al. 2020; Li et al. 2021), we utilize only the textual modality for experiments and adopt the weighted-F1 metric for evaluation. Additionally, we perform a paired t-test (Kim 2015) to assess the statistical significance of performance improvements.

Baseline Models and Implementation Details. To comprehensively evaluate DVL-CER, we compare it with state-of-the-art baselines, covering both single-view and multi-view CER methods (see Appendix B for a complete list). DialogueCRN serves as the backbone of the conversational context network in DVL-CER, except where noted otherwise. The reported results are averaged over 20 random runs on the test set to ensure reliable performance evaluation. Additional experimental details are available in Appendix C.

[1]Appendices are available at https://drive.google.com/file/d/1pRNxKGoDQVgNy6GGpiC6XgyJzRzoV00O/view

Method	IEMOCAP	EmoryNLP
Single-view methods		
DialogueGCN (Ghosal et al. 2019)	64.18	-
RGAT (Ishiwatari et al. 2020)	65.22	34.42
SKAIG (Li et al. 2021)	66.96	38.88
DialogXL (Shen et al. 2021a)	65.94	34.73
DAG-ERC-W (Shen et al. 2021b)	68.03	39.02
CoMPM (Lee and Lee 2022)	69.46	38.93
VAE (Ong et al. 2022)	68.23	-
CoG-BART (Li, Yan, and Qiu 2022)	66.18	39.04
SUNET (Song et al. 2023)	68.96	39.89
DAG-ERC-X (Quan et al. 2023)	68.50	39.19
denoiseGNN (Gan et al. 2024)	69.70	39.70
Multi-view methods		
DialogueRNN† (Majumder et al. 2019)	64.65	37.54
COSMIC (Ghosal et al. 2020)	65.28	38.11
DialogueCRN† (Hu, Wei, and Huai 2021)	67.53	38.79
MVN (Ma et al. 2022)	65.99	-
COSMIC+HCL (Yang et al. 2022)	66.23	38.96
HMVDM (Ruan et al. 2022)	67.96	38.46
SACL-LSTM (Hu et al. 2023)	69.22	39.65
DualRAN (Li, Wang, and Zeng 2024)	69.17	39.18
DialogueCRN‡	69.01	38.97
DVL-CER	**70.11** (↑ 1.10)	**39.92** (↑ 0.95)

Table 2: Performance comparison of various CER methods on IEMOCAP and EmoryNLP. †Results are taken from Hu et al. (2023). ‡Results are from our replication.

4.2 Comparison with State of the Art

Table 2 compares the performance of DVL-CER with the baselines. The results indicate that: (1) DVL-CER comparatively outperforms all baselines across both datasets, demonstrating its effectiveness in CER. By leveraging information from both utterance and emotion views, our approach generates high-quality emotion distributions, leading to more accurate emotion recognition. (2) Compared with the default DialogueCRN backbone, DVL-CER improves performance, suggesting that the dual-view projection heads effectively transfer valuable information from the emotion shift network to the conversational context network (addressing **Questions I** and **II**). Overall, DVL-CER achieves leading performance on both datasets, highlighting the advantages of our dual-view learning framework.

4.3 Applying DVL-CER to Different CER Baselines

The proposed DVL-CER is a flexible, context network-agnostic framework that can be integrated with various CER baseline models. To validate its effectiveness and generalizability, we apply DVL-CER to several prominent CER models: DialogueRNN, COSMIC, DAG-ERC-W, RGAT, and SACL-LSTM, which represent different mainstream ap-

Backbone	DVL-CER	IEMOCAP	EmoryNLP
DialogueRNN		66.03	37.73
DialogueRNN	✓	67.77* (↑ 1.74)	38.30* (↑ 0.57)
COSMIC		67.45	38.67
COSMIC	✓	68.96* (↑ 1.51)	39.20* (↑ 0.53)
RGAT		65.83	37.53
RGAT	✓	66.72* (↑ 0.89)	38.43* (↑ 0.90)
DAG-ERC-W		65.81	38.34
DAG-ERC-W	✓	66.22* (↑ 0.41)	39.20* (↑ 0.86)
SACL-LSTM		69.40	39.22
SACL-LSTM	✓	70.67* (↑ 1.27)	39.49 (↑ 0.27)

Table 3: Performance of the proposed DVL-CER framework when integrated with other CER baselines across two benchmarks. *Results show significant test p-value < 0.05 compared to their respective baseline models.

proaches to conversational context modeling. As shown in Table 3, DVL-CER consistently provides significant improvements over both single-view and multi-view baselines. This demonstrates that dual heterogeneous view learning enhances the accuracy and robustness of emotion distributions (addressing **Questions I** and **II**).

4.4 Comparison with Other Dual-View Learning Strategies

A key contribution of DVL-CER is its dual-view projection learning strategy. To assess whether the emotion view is beneficial across CER domains or specific to this strategy, we replace the dual-view projection learning strategy with other dual-view learning strategies, including dual-view feature aggregation learning, dual-view subspace learning, and dual-view alignment learning. These strategies integrate utterance and emotion views differently (see Appendix D for architectural details and Section 4.6 for theoretical analysis).

Our results on the IEMOCAP dataset, shown in the second column of Table 4, reveal that: (1) All dual-view learning strategies benefit from the emotion view, surpassing the DialogueCRN baseline and indicating successful emotion shift modeling (addressing **Question I**). (2) Among these strategies, our dual-view projection learning achieves the highest performance in DVL-CER. Similar results are noted on the EmoryNLP dataset, as detailed in Appendix.

4.5 Ablation Study

We conduct an ablation study on DVL-CER to gain insights into its behavior and performance. Table 5 presents the impact of various architectural components on model performance, and Table 6 explores the influence of different initial emotion settings and the application of the dual-view projection learning strategy to different interaction objects. In addition, Figure 3 shows the effect of varying emotion shift spans.

Detailed Ablation on Projectors. The ablation study on the two projectors, shown in the third row of Table 5, indicates that removing either projector significantly hinders view interaction, leading to a notable decline in performance.

The importance of ReLU in Projectors. Introducing the ReLU non-linear activation function in the projectors enhances the interaction between heterogeneous views, as shown in the fourth row of Table 5.

Impact of Supervision in Emotion View. The fifth row of Table 5 highlights the critical role of the loss $\mathcal{L}_{MSE}^{\psi}$ in model training. Without emotion-view supervision, the model's performance deteriorates, consistent with previous studies (Zhai et al. 2019; Henaff 2020).

Impact of Emotion Verification Chain. The sixth row of Table 5 indicates that verifying the emotion distributions predicted by the backbone substantially improves performance, indicating effective emotion connections in our emotion shift (addressing **Question I**).

Emotion Distribution versus Emotion Representation. Table 6 illustrates that emotion representations from two views are difficult to project and learn from each other, as discussed in Section 3.4.

Choice of Initial Emotions. To assess the role of initial emotions in emotion shift, we test two alternatives: a real but less informative one-hot label vector, and a zero vector that zeros out the initial emotion variable while maintaining consistency in the framework structure.

Results on the IEMOCAP dataset, as shown in Table 6, indicate that initializing with the emotion distribution is better than with the one-hot vector, and much better than with the zero vector (addressing **Question I**). Similar observations are noted on the EmoryNLP dataset, where the difference between the one-hot and zero vectors is minimal, primarily due to disruptions in the emotion verification chain.

Effect of Emotion Shift Span. We experiment with different emotion shift spans to explore the optimal threshold. As shown in Figure 3, the emotion shift modeling consistently provides valuable information to the context network across varying span settings (addressing **Question I**). The best performance for both datasets is achieved with an emotion shift span of $w = 4$, which is used as the default in our study.

4.6 Understanding the Learned Emotion Distributions

Finally, we explore the unique properties of our dual-view projection learning strategy compared to other dual-view learning strategies. As prerequisites, we discuss two important attributes: consistency and independence (Sun 2013). We use the similarity metric $\text{Sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v}/(\|\boldsymbol{u}\|\|\boldsymbol{v}\|)$ to evaluate the similarity among the two views and their projections. Consistency correlates positively with similarity, while independence correlates negatively. Theoretical and mathematical insights are as follows:

- **Dual-view feature aggregation learning** adds emotion distributions from both views for emotion prediction, maximizing independence for complementarity. Here, $\text{Sim}(e_\theta, e_\psi)$ shows the highest independence.
- **Dual-view subspace learning** use two projections to map both views into a common subspace where independence is filtered out and cross-view correlation is maximized. Hence, $\text{Sim}(e_\theta, e_\psi)$ before projections indicates

Strategy	IEMOCAP	$\text{Sim}(e_\theta, f_\psi(e_\psi))$	$\text{Sim}(f_\theta(e_\theta), e_\psi)$	$\text{Sim}(e_\theta, e_\psi)$	$\text{Sim}(f_\theta(e_\theta), f_\psi(e_\psi))$
Dual-view feature aggregation learning	69.73	×	×	0.6837	×
Dual-view subspace learning	69.55	0.0962	0.1918	0.8563	0.8803
Dual-view alignment learning	69.19	×	×	0.9667	×
Dual-view projection learning	70.11	0.9629	0.8858	0.7270	0.6435

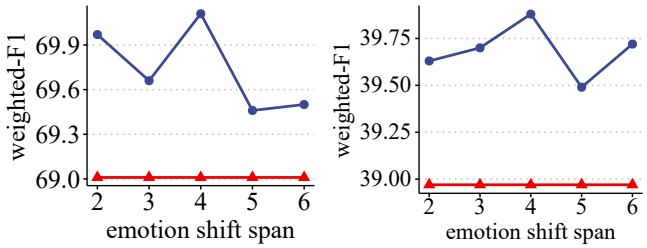
Table 4: Comparison of different dual-view learning strategies applied to DVL-CER on the IEMOCAP dataset. Symbol "×" indicates that the strategy cannot handle quantization.

Projector			Softmax		IEMOCAP	EmoryNLP	Comment
f_θ	f_ψ	ReLU	softmax_θ	softmax_ψ			
			✓		69.01	38.97	Loss: $\mathcal{L}_{CCM}$ (Baseline)
	✓	✓	✓	✓	69.70	39.68	Loss: $\mathcal{L}_{CCM} + \mathcal{L}_{ESM} + \lambda_2 \mathcal{L}^{\psi}_{MSE}$
✓		✓	✓	✓	69.30	39.42	Loss: $\mathcal{L}_{CCM} + \mathcal{L}_{ESM} + \lambda_1 \mathcal{L}^{\theta}_{MSE}$
✓	✓		✓	✓	69.48	39.66	Loss: $\mathcal{L}_{CCM} + \mathcal{L}_{ESM} + \lambda_1 \mathcal{L}^{\theta}_{MSE} + \lambda_2 \mathcal{L}^{\psi}_{MSE}$
✓	✓	✓	✓		69.63	39.65	Loss: $\mathcal{L}_{CCM} + \lambda_1 \mathcal{L}^{\theta}_{MSE} + \lambda_2 \mathcal{L}^{\psi}_{MSE}$
	✓	✓	✓		69.54	39.36	Loss: $\mathcal{L}_{CCM} + \lambda_2 \mathcal{L}^{\psi}_{MSE}$
✓	✓	✓	✓	✓	70.11	39.92	Loss: $\mathcal{L}_{CCM} + \mathcal{L}_{ESM} + \lambda_1 \mathcal{L}^{\theta}_{MSE} + \lambda_2 \mathcal{L}^{\psi}_{MSE}$

Table 5: Component-wise ablation study with Projectors and Softmax layers.

Interaction object	Initial emotion	IEMOCAP	EmoryNLP
representation r		68.47	38.50
distribution e		70.11	39.92
	zero vector	68.55	39.75
	one-hot vector	69.60	39.79
	e_θ	70.11	39.92

Table 6: Impact of dual-view projection learning strategy on interaction objects and initial emotions in emotion shifts.



(a) IEMOCAP (b) EmoryNLP

Figure 3: Effect of emotion shift span. The blue line: DVL-CER, and the red line: baseline model.

independence, while $\text{Sim}(f_\theta(e_\theta), f_\psi(e_\psi))$ after projections indicates consistency.

- **Dual-view alignment learning** aligns the emotion distributions of both views to maximize correlation, resulting in the highest consistency with $\text{Sim}(e_\theta, e_\psi)$.
- **Dual-view projection learning** transforms one view's emotion distribution to match the other's. Specifically, $\text{Sim}(f_\theta(e_\theta), e_\psi)$ and $\text{Sim}(e_\theta, f_\psi(e_\psi))$ measure the consistency between the two views in different view spaces, while $\text{Sim}(e_\theta, e_\psi)$ represents their independence.

Quantitative results on the IEMOCAP dataset, shown in Table 4, demonstrate that our dual-view projection learning strategy achieves a consistency level ($\text{Sim}(e_\theta, f_\psi(e_\psi))$) similar to dual-view alignment learning ($\text{Sim}(e_\theta, e_\psi)$), while preserving independence ($\text{Sim}(e_\theta, e_\psi)$) comparable to dual-view feature aggregation learning ($\text{Sim}(e_\theta, e_\psi)$). This better balance between consistency and independence across the utterance and emotion views makes our strategy outperform dual-view subspace learning and achieve state-of-the-art performance, as elaborated in Appendix E.

5 Conclusions

In this work, we present a dual-view learning framework that integrates utterance and emotion views to generate robust emotion distributions for CER. Through comprehensive empirical comparisons with existing CER and dual-view learning methods, we derive two main conclusions: (1) The emotion shift modeling approach is both effective and well-founded, as it successfully establishes connections between emotions. (2) The dual-view projection learning strategy effectively integrates information from multiple views, enhancing single-view learning process while maintaining a balance between cross-view consistency and view-specific independence. Future research could explore extending the emotion shift modeling to multi-modal CER. Additionally, the dual-view projection learning strategy shows potential for applications in multi-view supervised learning tasks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 62076092, the Guangdong Basic and Applied Basic Research Foundation under Grant No. 2024A1515010112, and the Changsha Science and Technology Bureau Foundation under Grant No. kq2402082. The authors express their sincere gratitude to the anonymous reviewers and the meta-reviewer for their insightful feedback and constructive suggestions on the paper.

References

Andrew, G.; Arora, R.; Bilmes, J.; and Livescu, K. 2013. Deep canonical correlation analysis. In *International conference on machine learning*, 1247–1255.

Babanejad, N.; Davoudi, H.; Agrawal, A.; An, A.; and Papagelis, M. 2024. The Role of Preprocessing for Word Representation Learning in Affective Tasks. *IEEE Transactions on Affective Computing*, 254–272.

Bansal, K.; Agarwal, H.; Joshi, A.; and Modi, A. 2022. Shapes of emotions: Multimodal emotion recognition in conversations via emotion shifts. In *Proceedings of the First Workshop on Performance and Interpretability Evaluations of Multimodal, Multipurpose, Massive-Scale Models*, 44–56.

Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Celikyilmaz, A.; and Choi, Y. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4762–4779.

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

Chatterjee, A.; Narahari, K. N.; Joshi, M.; and Agrawal, P. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th international workshop on semantic evaluation*, 39–48.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607.

Faruqui, M.; Dodge, J.; Jauhar, S. K.; Dyer, C.; Hovy, E. H.; and Smith, N. A. 2015. Retrofitting Word Vectors to Semantic Lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, 1606–1615.

Felbo, B.; Mislove, A.; Søgaard, A.; Rahwan, I.; and Lehmann, S. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1615–1625.

Gan, C.; Zheng, J.; Zhu, Q.; Jain, D. K.; and Štruc, V. 2024. A graph neural network with context filtering and feature correction for conversational emotion recognition. *Information Sciences*, 120017.

Gao, Q.; Cao, B.; Guan, X.; Gu, T.; Bao, X.; Wu, J.; Liu, B.; and Cao, J. 2022. Emotion recognition in conversations with emotion shift detection based on multi-task learning. *Knowledge-Based Systems*, 108861.

Ghosal, D.; Majumder, N.; Gelbukh, A. F.; Mihalcea, R.; and Poria, S. 2020. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2470–2481.

Ghosal, D.; Majumder, N.; Mihalcea, R.; and Poria, S. 2021. Exploring the role of context in utterance-level emotion, act and intent classification in conversations: An empirical study. In *Findings of the Association for Computational Linguistics: ACL 2021*, 1435–1449.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. F. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 154–164.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 1789–1819.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 21271–21284.

Henaff, O. 2020. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, 4182–4192.

Hu, D.; Bao, Y.; Wei, L.; Zhou, W.; and Hu, S. 2023. Supervised Adversarial Contrastive Learning for Emotion Recognition in Conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 10835–10852.

Hu, D.; Wei, L.; and Huai, X. 2021. DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 7042–7052.

Ishiwatari, T.; Yasuda, Y.; Miyazaki, T.; and Goto, J. 2020. Relation-aware Graph Attention Networks with Relational Position Encodings for Emotion Recognition in Conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 7360–7370.

Kim, T. K. 2015. T test as a parametric statistic. *Korean journal of anesthesiology*, 540–546.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kumar, A.; Dogra, P.; and Dabas, V. 2015. Emotion analysis of Twitter using opinion mining. In *Proceedings of the 8th International Conference on Contemporary Computing*.

Lee, J.; and Lee, W. 2022. CoMPM: Context Modeling with Speaker's Pre-trained Memory Tracking for Emotion Recognition in Conversation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, 5669–5679.

Li, J.; Lin, Z.; Fu, P.; and Wang, W. 2021. Past, Present, and Future: Conversational Emotion Recognition through Structural Modeling of Psychological Knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 1204–1214.

Li, J.; Wang, X.; Liu, Y.; and Zeng, Z. 2024. CFN-ESA: A Cross-Modal Fusion Network With Emotion-Shift Awareness for Dialogue Emotion Recognition. *IEEE Transactions on Affective Computing*, 1–16.

Li, J.; Wang, X.; and Zeng, Z. 2024. A dual-stream recurrence-attention network with global–local awareness for emotion recognition in textual dialog. *Engineering Applications of Artificial Intelligence*, 107530.

Li, S.; Yan, H.; and Qiu, X. 2022. Contrast and generation make bart a good dialogue emotion recognizer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11002–11010.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

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

Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A. F.; and Cambria, E. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6818–6825.

Meng, T.; Zhang, F.; Shou, Y.; Shao, H.; Ai, W.; and Li, K. 2024. Masked Graph Learning with Recurrent Alignment for Multimodal Emotion Recognition in Conversation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 1–14.

Mohammad, S. M. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20, 000 English Words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 174–184.

Nabi, R. L.; and Green, M. C. 2015. The role of a narrative's emotional flow in promoting persuasive outcomes. *Media Psychology*, 137–162.

Ong, D.; Su, J.; Chen, B.; Luu, A. T.; Narendranath, A.; Li, Y.; Sun, S.; Lin, Y.; and Wang, H. 2022. Is discourse role important for emotion recognition in conversation? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11121–11129.

Poria, S.; Hazarika, D.; Majumder, N.; and Mihalcea, R. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE transactions on affective computing*, 108–132.

Quan, X.; Wu, S.; Chen, J.; Shen, W.; and Yu, J. 2023. Multi-Party Conversation Modeling for Emotion Recognition. *IEEE Transactions on Affective Computing*, 1–17.

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

Ruan, Y.-P.; Zheng, S.-K.; Li, T.; Wang, F.; and Pei, G. 2022. Hierarchical and multi-view dependency modelling network for conversational emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 7032–7036.

Shen, W.; Chen, J.; Quan, X.; and Xie, Z. 2021a. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 13789–13797.

Shen, W.; Wu, S.; Yang, Y.; and Quan, X. 2021b. Directed Acyclic Graph Network for Conversational Emotion Recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 1551–1560.

Song, R.; Giunchiglia, F.; Shi, L.; Shen, Q.; and Xu, H. 2023. SUNET: Speaker-utterance interaction Graph Neural Network for Emotion Recognition in Conversations. *Engineering Applications of Artificial Intelligence*, 106315.

Sun, S. 2013. A survey of multi-view machine learning. *Neural computing and applications*, 2031–2038.

Xue, Z.; Du, J.; Du, D.; and Lyu, S. 2019. Deep low-rank subspace ensemble for multi-view clustering. *Information Sciences*, 210–227.

Yan, X.; Hu, S.; Mao, Y.; Ye, Y.; and Yu, H. 2021. Deep multi-view learning methods: A review. *Neurocomputing*, 106–129.

Yang, L.; Shen, Y.; Mao, Y.; and Cai, L. 2022. Hybrid curriculum learning for emotion recognition in conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11595–11603.

Yang, Z.-X.; Tang, L.; Zhang, K.; and Wong, P. K. 2018. Multi-view CNN feature aggregation with ELM autoencoder for 3D shape recognition. *Cognitive Computation*, 908–921.

Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 5634–5641.

Zahiri, S. M.; and Choi, J. D. 2018. Emotion Detection on TV Show Transcripts with Sequence-Based Convolutional Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 44–52.

Zhai, X.; Oliver, A.; Kolesnikov, A.; and Beyer, L. 2019. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1476–1485.

Zhao, W.; Zhao, Y.; and Lu, X. 2022. CauAIN: Causal Aware Interaction Network for Emotion Recognition in Conversations. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, 4524–4530.

Zhong, P.; Wang, D.; and Miao, C. 2019. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 165–176.