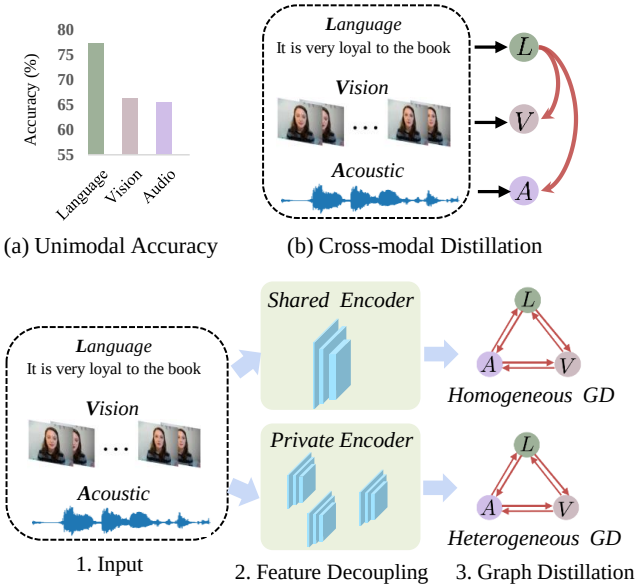
Decoupled Multimodal Distilling for Emotion Recognition

Yong Li, Yuanzhi Wang, Zhen Cui*
PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China.
{yong.li, yuanzhiwang, zhen.cui}@njust.edu.cn

Abstract

Human multimodal emotion recognition (MER) aims to perceive human emotions via language, visual and acoustic modalities. Despite the impressive performance of previous MER approaches, the inherent multimodal heterogeneities still haunt and the contribution of different modalities varies significantly. In this work, we mitigate this issue by proposing a decoupled multimodal distillation (DMD) approach that facilitates flexible and adaptive crossmodal knowledge distillation, aiming to enhance the discriminative features of each modality. Specially, the representation of each modality is decoupled into two parts, i.e., modality-irrelevant/-exclusive spaces, in a self-regression manner. DMD utilizes a graph distillation unit (GD-Unit) for each decoupled part so that each GD can be performed in a more specialized and effective manner. A GD-Unit consists of a dynamic graph where each vertice represents a modality and each edge indicates a dynamic knowledge distillation. Such GD paradigm provides a flexible knowledge transfer manner where the distillation weights can be automatically learned, thus enabling diverse crossmodal knowledge transfer patterns. Experimental results show DMD consistently obtains superior performance than state-of-the-art MER methods. Visualization results show the graph edges in DMD exhibit meaningful distributional patterns w.r.t. the modality-irrelevant/-exclusive feature spaces. Codes are released at https://github.com/mdswyz/DMD.



(a) Unimodal Accuracy (b) Cross-modal Distillation

(c) Our proposed Decoupled Multimodal Distillation

Figure 1. (a) illustrates the significant emotion recognition discrepancies using unimodality, adapted from Mult [28]. (b) shows the conventional cross-modal distillation. (c) shows our proposed decoupled multimodal distillation (DMD) method. DMD consists of two graph distillation (GD) units: homogeneous GD and heterogeneous GD. The decoupled GD paradigm decreases the burden of absorbing knowledge from the heterogeneous data and allows each GD to be performed in a more specialized and effective manner.

1. Introduction

Human multimodal emotion recognition (MER) aims to perceive the sentiment attitude of humans from video clips [13, 17]. The video flows involve time-series data from various modalities, e.g., language, acoustic, and vision. This rich multimodality facilitates us in understanding human behaviors and intents from a collaborative perspective. Recently, MER has become one of the most active research topics of affective computing with abundant appealing applications, such as intelligent tutoring systems [24], product feedback estimation [18], and robotics [15].

For MER, different modalities in the same video segment are often complementary to each other, providing extra cues for semantic and emotional disambiguation. The core part of MER is multimodal representation learning and fusion, in which a model aims to encode and integrate representations from multiple modalities to understand the emotion behind the raw data. Despite the achievement of the

* The corresponding author.

mainstream MER methods [7, 28, 33], the intrinsic heterogeneities among different modalities still perplex us and increase the difficulty of robust multimodal representation learning. Different modalities, e.g., image, language, and acoustic, contain different ways of conveying semantic information. Typically, the language modality consists of limited transcribed texts and has more abstract semantics than nonverbal behaviors. As illustrated in Fig. 1 (a), language plays the most important role in MER and the intrinsic heterogeneities result in significant performance discrepancies among different modalities [25, 28, 31].

One way to mitigate the conspicuous modality heterogeneities is to distill the reliable and generalizable knowledge from the strong modality to the weak modality [6], as illustrated in Fig. 1 (b). However, such manual assignment for the distillation direction or weights should be cumbersome because there are various potential combinations. Instead, the model should learn to automatically adapt the distillation according to different examples, e.g, many emotions are easier to recognize via language while some are easier by vision. Furthermore, the significant feature distribution mismatch cross the modalities makes the direct crossmodal distillation sub-optimal [21, 37].

To this end, we propose a decoupled multimodal distillation (DMD) method to learn dynamic distillations across modalities, as illustrated in Fig. 1 (c). Typically, the features of each modality are decoupled into modality-irrelevant/-exclusive spaces via shared encoder and the private encodes, respectively. As to achieve the feature decoupling, we devise a self-regression mechanism that predicts the decoupled modality features and then regresses them self-supervisedly. To consolidate the feature decoupling, we incorporate a margin loss that regularizes the proximity in relationships of the representations across modalities and emotions. Consequently, the decoupled GD paradigm would decrease the burden of absorbing knowledge from the heterogeneous data and allows each GD to be performed in a more specialized and effective manner.

Based on the decoupled multimodal feature spaces, DMD utilizes a graph distillation unit (GD-Unit) in each space so that the crossmodal knowledge distillation can be performed in a more specialized and effective manner. A GD-Unit consists of a graph that (1) vertices representing the representations or logits from the modalities and (2) edges indicating the knowledge distillation directions and weights. As the distribution gap among the modality-irrelevant (homogeneous) features is sufficiently reduced, GD can be directly applied to capture the inter-modality semantic correlations. For the modality-exclusive (heterogeneous) counterparts, we exploit the multimodal transformer [28] to build the semantic alignment and bridge the distribution gap. The cross-modal attention mechanism in the multimodal transformer reinforces the multimodal representations and reduces the discrepancy between the high-level semantic concepts that exist in different modalities. For simplification, we respectively name the distillation on the decoupled multimodal features as homogeneous graph knowledge distillation (HomoGD) and heterogeneous graph knowledge distillation (HeteroGD). The reformulation allows us to explicitly explore the interaction between different modalities in each decoupled space.

The contributions of this work can be summarized as:

- We propose a decoupled multimodal distillation framework, Decoupled Multimodal Distillation (DMD), to learn the dynamic distillations across modalities for robust MER. In DMD, we explicitly decouple the multimodal representations into modality-irrelevant/-exclusive spaces to facilitate KD on the two decoupled spaces. DMD provides a flexible knowledge transfer manner where the distillation directions and weights can be automatically learned, thus enabling flexible knowledge transfer patterns.

- We conduct comprehensive experiments on public MER datasets and obtain superior or comparable results than the state-of-the-arts. Visualization results verify the feasibility of DMD and the graph edges exhibit meaningful distributional patterns w.r.t. HomoGD and HeteroGD.

2. Related Works

2.1. Multimodal emotion recognition

Multimodal emotion recognition (MER) aims to infer human sentiment from the language, visual and acoustic information embedded in the video clips. The heterogeneity across modalities can provide various levels of information for MER. The mainstream MER approaches can be divided into two categories: fusion strategy-based [14, 33, 34] and crossmodal attention-based [13, 17, 28].

The former aims to design sophisticated multimodal fusion strategies to generate discriminative multimodal representations, e.g., Zadeh *et al.* [33] designed a Tensor Fusion Network (TFN) that can fuse multimodal information progressively. However, the inherent heterogeneity and the intrinsic information redundancy across modalities hinders the fusion between the multimodal features. Therefore, some work aims to explore the characteristics and commonalities of multimodal representations via feature decoupling to facilitate more effective multimodal representation fusion [7, 29, 32]. Hazarika *et al.* [7] decomposed multimodal features into modality-invariant/-specific components to learn the refined multimodal representations. The decoupled multimodal representations reduce the information redundancy and provide a holistic view of the multimodal data. Recently, crossmodal attention-based

approaches have driven the development of MER since they learn the cross-modal correlations to obtain the reinforced modality representation. A representative work is MulT [28]. This work proposes the multimodal transformer that consists of a cross-modal attention mechanism to learn the potential adaption and correlations from one modality to another, thereby achieving semantic alignment between modalities. Lv *et al.* [17] designed a progressive modality reinforcement method based on [28], it aims to learn the potential adaption and correlations from multimodal representation to unimodal representation. Our proposed DMD has an essential difference with the previous feature-decoupling methods [7, 29, 32] because DMD is capable of distilling cross-modal knowledge in the decoupled feature spaces.

2.2. Knowledge distillation

The concept of knowledge distillation (KD) was first proposed in [9] to transfer knowledge via minimizing the KL-Divergence between prediction logits of teachers and students. Subsequently, various KD methods were proposed [4, 5, 20, 39] based on [9] and further extended to distillation between intermediate features [1, 8, 22, 27].

Most KD methods focus on transferring knowledge from the teacher to the student, while some recent studies have used graph structures to explore the effective message passing mechanism between multiple teachers and students with multiple instances of knowledge [16, 19, 38]. Zhang *et al.* [38] proposed a graph distillation (GD) method for video classification, where each vertex represented a self-supervised teacher and edges represented the direction of distillation from multiple self-supervised teachers to the student. Luo *et al.* [16] considered the modality discrepancy to incorporate privileged information from the source domain and modeled a directed graph to explore the relationship between different modalities. Each vertex represented a modality and the edges indicated the connection strength (i.e., distillation strength) between one modality and another. Different from them, we aim to use exclusive GD-Units in the decoupled feature spaces to facilitate effective cross-modality distillation.

3. The Proposed Method

The framework of our DMD is illustrated in Fig. 2. It mainly consists of three parts: **multimodal feature decoupling**, **homogeneous GD** (HomoGD), **hetergeneous GD** (HeteroGD). Considering the significant distribution mismatch of modalities, we decouple multimodal representations into homogeneous and heterogeneous multimodal features through learning shared and exclusive multimodal encoders. The decoupling detail is introduced in Sec. 3.1. To facilitate a flexible knowledge transfer, we next distill the knowledge from homo/heterogeneous features, which are framed in two graph distillation units (GD-Unit), i.e., HomoGD and HeteroGD. In HomoGD, homogeneous multimodal features are mutually distilled to compensate the representation ability for each other. In HeteroGD, multimodal transformers are introduced to explicitly build inter-modal correlations and semantic alignment for further distilling. The GD detail is introduced in Sec. 3.2. Finally, the refined multimodal features through distilling are adaptively fused for robust MER. Below, we present the details of the three parts of DMD.

3.1. Multimodal feature decoupling

We consider three modalities, i.e., *language* (L), *visual* (V), *acoustic* (A). Firstly, we exploit three separate 1D temporal convolutional layers to aggregate temporal information and obtain the low-level multimodal features: $\widetilde{\mathbf{X}}_m \in \mathbb{R}^{T_m \times d_m}$, where $m \in \{L, V, A\}$ indicates a modality. After this shallow encoding, each modality preserves the input temporal dimension to facilitate handling unaligned and aligned cases simultaneously. Moreover, all modalities are scaled to the same feature dimension, i.e., $d_L = d_V = d_A = d$, for convenient subsequent feature decoupling.

To decouple the multimodal features into homogeneous (modality-irrelevant) part $\mathbf{X}_m^{\text{com}}$ and heterogeneous (modality-exclusive) part $\mathbf{X}_m^{\text{prt}}$, we exploit a shared multimodal encoder $\mathcal{E}^{\text{com}}$ and three private encoders $\mathcal{E}_m^{\text{prt}}$ to explicitly predict the decoupled features. Formally,

$$\mathbf{X}_m^{\text{com}} = \mathcal{E}^{\text{com}}(\widetilde{\mathbf{X}}_m), \mathbf{X}_m^{\text{prt}} = \mathcal{E}_m^{\text{prt}}(\widetilde{\mathbf{X}}_m). \tag{1}$$

To distinguish the differences between $\mathbf{X}_m^{\text{com}}$ and $\mathbf{X}_m^{\text{prt}}$ and mitigate the feature ambiguity, we synthesize the vanilla coupled features $\widetilde{\mathbf{X}}_m$ in a self-regression manner. Mathematically speaking, we concatenate $\mathbf{X}_m^{\text{com}}$ and $\mathbf{X}_m^{\text{prt}}$ for each modality and exploit a private decoder $\mathcal{D}_m$ to produce the coupled feature, i.e., $\mathcal{D}_m([\mathbf{X}_m^{\text{com}}, \mathbf{X}_m^{\text{prt}}])$. Subsequently, the coupled feature will be re-encoded via the private encoders $\mathcal{E}_m^{\text{prt}}$ to regress the heterogeneous features. The notation [.] means feature concatenation. Formally, the discrepancy between the vanilla/synthesized coupled multimodal features can be formulated as:

$$\mathcal{L}_{\text{rec}} = \|\widetilde{\mathbf{X}}_m - \mathcal{D}_m([\mathbf{X}_m^{\text{com}}, \mathbf{X}_m^{\text{prt}}])\|_F^2. \tag{2}$$

Further, the discrepancy between the vanilla/synthesized heterogeneous features can be formulated as:

$$\mathcal{L}_{\text{cyc}} = \|\mathbf{X}_m^{\text{prt}} - \mathcal{E}_m^{\text{prt}}(\mathcal{D}_m([\mathbf{X}_m^{\text{com}}, \mathbf{X}_m^{\text{prt}}]))\|_F^2. \tag{3}$$

For the above reconstruction losses, it still cannot guarantee the complete feature decoupling. In fact, information can freely leak between representations, e.g., all the modality information can be merely encoded in $\mathbf{X}_m^{\text{prt}}$ so that the decoders can easily synthesize the input, leaving homogeneous multimodal features meaningless. To consolidate the

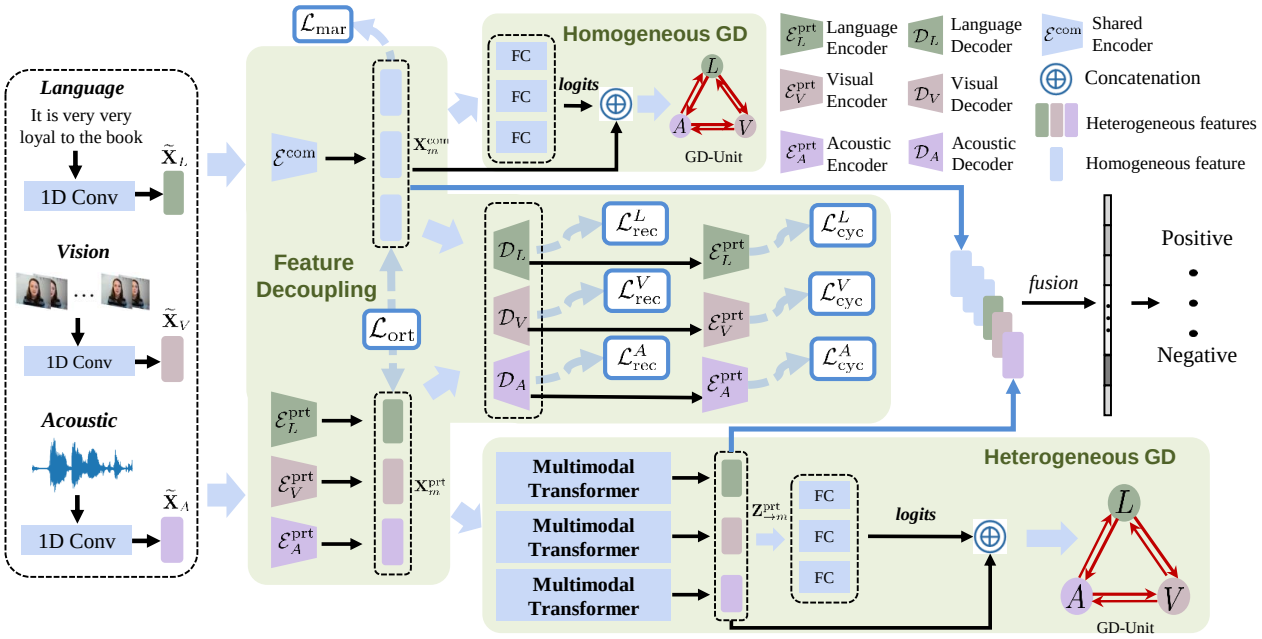


Figure 2. The framework of DMD. Given the input multimodal data, DMD encodes their respective shallow features $\widetilde{\mathbf{X}}_m$, where $m \in \{L, V, A\}$. In *feature decoupling*, DMD exploits the decoupled homo-/heterogeneous multimodal features $\mathbf{X}_m^{\text{com}}$ / $\mathbf{X}_m^{\text{prt}}$ via the shared and exclusive encoders, respectively. $\mathbf{X}_m^{\text{prt}}$ will be reconstructed in a self-regression manner (Sec. 3.1). Subsequently, $\mathbf{X}_m^{\text{com}}$ will be fed into a GD-Unit for adaptive knowledge distillation in *HomoGD*. In *HeteroGD*, $\mathbf{X}_m^{\text{prt}}$ are reinforced to $\mathbf{Z}_{\to m}^{\text{prt}}$ via multimodal transformers to bridge the distribution gap. The GD-Unit in *HeteroGD* takes $\mathbf{Z}_{\to m}^{\text{prt}}$ as input for distillation (Sec. 3.2). Finally, $\mathbf{X}_m^{\text{com}}$ and $\mathbf{Z}_{\to m}^{\text{prt}}$ will be adaptively fused for MER.

feature decoupling, we argue that homogeneous representations from the same emotion but different modalities should be more similar than those from the same modality but different emotions. To this end, we define a margin loss as:

$$\mathcal{L}_{\text{mar}} = \frac{1}{|S|}\sum_{(i,j,k)\in S} \max(0, \alpha - \cos(\mathbf{X}_{m[i]}^{\text{com}}, \mathbf{X}_{m[j]}^{\text{com}}) + \cos(\mathbf{X}_{m[i]}^{\text{com}}, \mathbf{X}_{m[k]}^{\text{com}})), \tag{4}$$

where we collect a triple tuple set $S = \{(i,j,k)|m[i] \neq m[j], m[i] = m[k], c[i] = c[j], c[i] \neq c[k]\}$. The $m[i]$ is the modality of sample i, $c[i]$ is the class label of sample i, and $\cos(\cdot,\cdot)$ means the cosine similarity between two feature vectors. The loss in Eq. 4 constrains the homogeneous features that belong to the same emotion but different modalities or vice versa to differ, and thereby avoids deriving trivial homogeneous features. α is a distance margin. The distances of positive samples (same *emotion*; different *modalities*) are constrained to be smaller than that of negative samples (same *modality*; different *emotions*) by the margin α. Considering that the decoupled features respectively capture the modality-irrelevant/-exclusive characteristics, we further formulate a soft orthogonality to reduce the information redundancy between the homogeneous and the heterogeneous multimodal features:

$$\mathcal{L}_{\text{ort}} = \sum_{m\in\{L,V,A\}} \cos(\mathbf{X}_m^{\text{com}}, \mathbf{X}_m^{\text{prt}}). \tag{5}$$

Finally, we combine these constraints to form the decoupling loss,

$$\mathcal{L}_{\text{dec}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{cyc}} + \gamma(\mathcal{L}_{\text{mar}} + \mathcal{L}_{\text{ort}}), \tag{6}$$

where γ is the balance factor.

3.2. GD with Decoupled Multimodal Features

For the decoupled homogeneous and heterogeneous multimodal features, we design a graph distillation unit (GD-Unit) on each of them to conduct adaptive knowledge distillation. Typically, a GD-Unit consists of a directed graph $\mathcal{G}$. Let v_i denote a node w.r.t a modality and $w_{i\to j}$ indicates the distillation strength from modality v_i to v_j. The distillation from v_i to v_j is defined as the difference between their corresponding logits, denoted with $\epsilon_{i\to j}$. Let $\mathbf{E}$ denotes the distillation matrix with $E_{ij} = \epsilon_{i\to j}$. For a target modality j, the weighted distillation loss can be formulated by considering the injection edges as,

$$\zeta_{:j} = \sum_{v_i\in\mathcal{N}(v_j)} w_{i\to j} \times \epsilon_{i\to j}, \tag{7}$$

where $\mathcal{N}(v_j)$ represents the set of vertices injected to v_j.

To learn a dynamic and adaptive weight that corresponds to the distillation strength w, we propose to encode the modality logits and the representations into the graph edges. Formally,

$$w_{i\to j} = g([[f(\mathbf{X}_i, \theta_1), \mathbf{X}_i], [f(\mathbf{X}_j, \theta_1), \mathbf{X}_j]], \theta_2), \quad (8)$$

where $[\cdot,\cdot]$ means feature concatenation, g is a fully-connected (FC) layer with the learnable parameters θ_2, and f is a FC layer for regressing logits with the parameters θ_1. The graph edge weights $\mathbf{W}$ with $W_{ij} = w_{i\to j}$ can be constructed and learned by repetitively applying Eq. 8 over all pairs of modalities. To reduce the scale effects, we normalize $\mathbf{W}$ through the $softmax$ operation. Consequently, the graph distillation loss to all modalities can be written as:

$$\mathcal{L}_{\text{dtl}} = \|\mathbf{W} \odot \mathbf{E}\|_1, \quad (9)$$

where $\odot$ means element-wise product. Obviously, the distillation graph in a GD-Unit provides a base for learning dynamic inter-modality interactions. Meanwhile, it facilitates a flexible knowledge transfer manner where the distillation strengths can be automatically learned, thus enabling diverse knowledge transfer patterns. Below, we elaborate on the details of HomoGD and HeteroGD.

HomoGD. As illustrated in Fig. 2, for the decoupled homogeneous features $\mathbf{X}_m^{\text{com}}$, as the distribution gap among the modalities is already reduced sufficiently, we input the features $\mathbf{X}_m^{\text{com}}$ and the corresponding logits $f(\mathbf{X}_m^{\text{com}})$ to a GD-Unit and calculate the graph edge matrix $\mathbf{W}$ and the distillation loss matrix $\mathbf{E}$ according to Eq. 8. Then, the overall distillation loss $\mathcal{L}_{\text{dtl}}^{\text{homo}}$ can be naturally obtained via Eq. 9.

HeteroGD. The decoupled heterogeneous features $\mathbf{X}_m^{\text{prt}}$ focus on the diversity and the unique characteristics of each modality, and thus exhibit a significant distribution gap. To mitigate this issue, we exploit the multimodal transformer [28] to bridge the feature distribution gap and build the modality adaptation. The core of the multimodal transformer is the crossmodal attention unit (CA), which receives features from a pair of modalities and fuses cross-modal information. Take the language modality $\mathbf{X}_L^{\text{prt}}$ as the source and the visual modality $\mathbf{X}_V^{\text{prt}}$ as the target, the cross-modal attention can be defined as: $\mathbf{Q}_V = \mathbf{X}_V^{\text{prt}}\mathbf{P}_q$, $\mathbf{K}_L = \mathbf{X}_L^{\text{prt}}\mathbf{P}_k$, and $\mathbf{V}_L = \mathbf{X}_L^{\text{prt}}\mathbf{P}_v$ where $\mathbf{P}_q, \mathbf{P}_k, \mathbf{P}_v$ are the learnable parameters. The individual head of can be expressed as:

$$\mathbf{Z}_{L\to V}^{\text{prt}} = \text{softmax}(\frac{\mathbf{Q}_V\mathbf{K}_L^\top}{\sqrt{d}})\mathbf{V}_L, \quad (10)$$

where $\mathbf{Z}_{L\to V}^{\text{prt}}$ is the enhanced features from Language to Visual, d means the dimension of $\mathbf{Q}_V$ and $\mathbf{K}_L$. For the three modalities in MER, each modality will be reinforced by the two others and the resulting features will be concatenated. For each target modality, we concatenate all enhanced features from other modalities to the target as the reinforced features, denotes with $\mathbf{Z}_{\to m}^{\text{prt}}$, which are used in the distillation loss function as $\mathcal{L}_{\text{dtl}}^{\text{hetero}}$ that can be naturally obtained via Eq. 9.

Feature fusion. We use the reinforced heterogeneous features $\mathbf{Z}_{\to m}^{\text{prt}}$ and the vanilla decoupled homogeneous features $\mathbf{X}_m^{\text{com}}$ for adaptive feature fusion with an adaptive weight learned from each of them. Hereby, we obtain the fused feature for multimodal emotion recognition.

3.3. Objective optimization

We integrate the above losses to reach the full objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_1\mathcal{L}_{\text{dec}} + \lambda_2\mathcal{L}_{\text{dtl}}, \quad (11)$$

where $\mathcal{L}_{\text{task}}$ is the emotion task related loss (here mean absolute error), $\mathcal{L}_{\text{dtl}} = \mathcal{L}_{\text{dtl}}^{\text{homo}} + \mathcal{L}_{\text{dtl}}^{\text{hetero}}$ means the distillation losses generated by HomoGD and HeteroGD, and λ_1, λ_2 control the importance of different constraints.

4. Experiments

Datasets. We evaluate DMD on CMU-MOSI [35] and CMU-MOSEI [36] datasets. The experiments are conducted under the word-aligned and unaligned settings for a more comprehensive comparison. **CMU-MOSI** consists of 2,199 short monologue video clips. The acoustic and visual features in CMU-MOSI are extracted at a sampling rate of 12.5 and 15 Hz, respectively. Among the samples, 1,284, 229 and 686 samples are used as training, validation and testing set. **CMU-MOSEI** contains 22,856 samples of movie review video clips from YouTube (approximately $10\times$ the size of CMU-MOSI). The acoustic and visual features were extracted at a sampling rate of 20 and 15 Hz, respectively. According to the predetermined protocol, 16,326 samples are used for training, the remainng 1,871 and 4,659 samples are used for for validation and testing. Each sample in CMU-MOSI and CMU-MOSEI was labeled with a sentiment score which ranges from -3 to 3, including *highly negative*, *negative*, *weakly negative*, *neutral*, *weakly positive*, *positive*, and *highly positive*. Following previous work [13, 17], we evaluate the MER performance using the following metrics: 7-class accuracy (ACC_7), binary accuracy (ACC_2) and F1 score.

Implementation details. On the two datasets, we extract the unimodal language features via GloVe [23] and obtain 300-dimensional word features. For a fair comparison with [7, 32] under the aligned setting, we additationally exploit a BERT-base-uncased pre-trained model [10] to obtain a 768-dimensional hidden state as the word features. For visual modality, each video frame was encoded via Facet [2] to represent the presence of the totally 35 facial action units [11, 12]. The acoustic modality was pro-

Table 1. Comparison on CMU-MOSI dataset. **Bold** is the best.

Methods	Setting	ACC_7 (%)	ACC_2 (%)	F1 (%)
EF-LSTM	Aligned	33.7	75.3	75.2
LF-LSTM		35.3	76.8	76.7
TFN [33]		32.1	73.9	73.4
LMF [14]		32.8	76.4	75.7
MFM [29]		36.2	78.1	78.1
RAVEN [30]		33.2	78.0	76.6
MCTN [26]		35.6	79.3	79.1
MulT [28]		40.0	83.0	82.8
PMR [17]		40.6	83.6	83.4
DMD (**Ours**)		**41.4**	**84.5**	**84.4**
MISA [7]*	Aligned	42.3	83.4	83.6
FDMER [32]*		44.1	84.6	84.7
DMD (**Ours**)*		**45.6**	**86.0**	**86.0**
EF-LSTM	Unaligned	31.0	73.6	74.5
LF-LSTM		33.7	77.6	77.8
RAVEN [30]		31.7	72.7	73.1
MCTN [26]		32.7	75.9	76.4
MulT [28]		39.1	81.1	81.0
PMR [17]		40.6	82.4	82.1
MICA [13]		40.8	82.6	82.7
DMD (**Ours**)		**41.9**	**83.5**	**83.5**

* means the input language features are BERT-based.

Table 2. Comparison on CMU-MOSEI dataset. **Bold** is the best.

Methods	Setting	ACC_7 (%)	ACC_2 (%)	F1 (%)
EF-LSTM	Aligned	47.4	78.2	77.9
LF-LSTM		48.8	80.6	80.6
Graph-MFN [36]		45.0	76.9	77.0
RAVEN [30]		50.0	79.1	79.5
MCTN [26]		49.6	79.8	80.6
MulT [28]		51.8	82.5	82.3
PMR [17]		52.5	83.3	82.6
DMD (**Ours**)		**53.7**	**85.0**	**84.9**
MISA [7]*	Aligned	52.2	85.5	85.3
FDMER [32]*		54.1	86.1	85.8
DMD (**Ours**)*		**54.5**	**86.6**	**86.6**
EF-LSTM	Unaligned	46.3	76.1	75.9
LF-LSTM		48.8	77.5	78.2
RAVEN [30]		45.5	75.4	75.7
MCTN [26]		48.2	79.3	79.7
MulT [28]		50.7	81.6	81.6
PMR [17]		51.8	83.1	82.8
MICA [13]		52.4	83.7	83.3
DMD (**Ours**)		**54.6**	**84.8**	**84.7**

* means the input language features are BERT-based.

cessed by COVAREP [3] to obtain the 74-dimensional features. *The detailed neural network configurations in DMD are listed in the supplementary file.* The optimal setting for λ_1, λ_2, γ was set as 0.1, 0.05, 0.1 via the MER performance on the validation set. We implemented all the experiments using PyTorch on a RTX 3090 GPU with 24GB memory. We set the training batch size as 16 and trained DMD for 30 epoches until convergence.

Table 3. Ablation study of the key components in DMD.

Dataset	FD	HomoGD	CA	HeteroGD	ACC_7	F1
MOSI	✓	✓	✓	✓	**41.9**	**83.5**
	✓	✓	✓	×	38.8	81.1
	✓	✓	×	✓	37.5	80.6
	✓	✓	×	×	37.2	80.8
	✓	×	×	×	34.7	79.3
	×	×	×	×	32.4	79.0
MOSEI	✓	✓	✓	✓	**54.6**	**84.7**
	✓	✓	✓	×	53.2	84.1
	✓	✓	×	✓	52.4	83.8
	✓	✓	×	×	52.4	84.3
	✓	×	×	×	51.6	82.8
	×	×	×	×	50.0	81.9

Table 4. Unimodal accuracy comparison on MOSEI dataset.

Methods	w/o FD	w/ FD
	Acc_2 (%) / F1 (%)	Acc_2 (%) / F1 (%)
L only	81.2 / 81.4	**82.7 / 82.5**
V only	58.2 / 52.2	**62.8 / 60.0**
A only	53.4 / 54.0	**64.9 / 62.5**
Mean	64.3 / 62.5	**70.1 / 68.3**
STD	12.1 / 13.4	**8.9 / 10.1**

Table 5. Ablation study of graph distillation (GD) on MulT.

Methods	CMU-MOSI			CMU-MOSEI		
	ACC_7	ACC_2	F1	ACC_7	ACC_2	F1
MulT	39.1	81.1	81.0	50.7	81.6	81.6
MulT (*w/ GD*)	39.4	82.2	82.2	51.0	82.3	82.5
DMD (**Ours**)	**41.9**	**83.5**	**83.5**	**54.6**	**84.8**	**84.7**

4.1. Comparison with the state-of-the-art

We compare DMD with the current state-of-the-art MER methods under the same dataset settings (unaligned or aligned), including EF-LSTM, LF-LSTM, TFN [33], LMF [14], MFM [29], RAVEN [30], Graph-MFN [36], MCTN [26], MulT [28], PMR [17], MICA [13], MISA [7], and FDMER [32].

Tab. 1 and Tab. 2 illustrate the comparison on CMU-MOSI and CMU-MOSEI datasets, respectively. Obviously, our proposed DMD obtains superior MER accuracy than other MER approaches under the unaligned and aligned settings. Compared with the feature-disentangling-based MER methods [7, 29, 32], our proposed DMD obtains consistent improvements, indicating the feasibility of the incorporated GD-Unit, which is capable of perceiving the various inter-modality dynamics. For a further inverstigation, we will visualize the learned graph edges in each GD-Unit in Sec. 4.2. DMD consistently outperforms the methods [13, 17, 28] that use multimodal transformer to learn crossmodal interactions and perform multimodal fusion. The reasons are two-fold: (1) DMD takes the modality-irrelevant/-exclusive spaces into consideration concurrently and recudes the information redundancy via feature decoupling. (2) DMD exploits twin GD-Units to adaptively distil knowledge among the modalities. On CMU-MOSEI dataset, Graph-MFN [36]

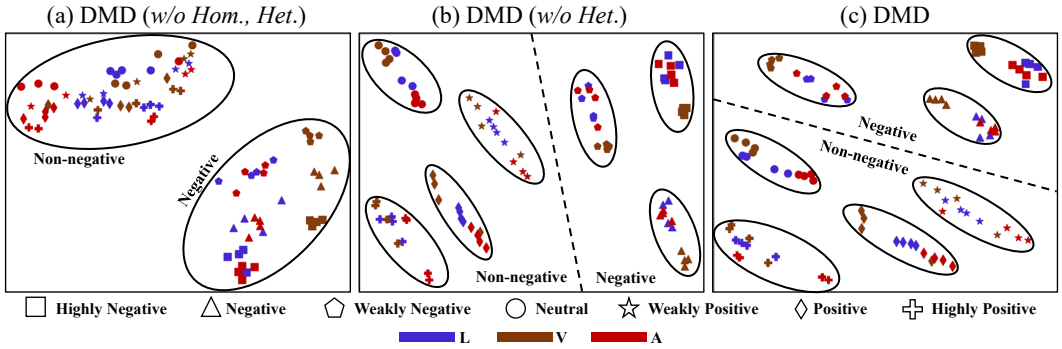


Figure 3. t-SNE visualization of decoupled homogeneous space on MOSEI. DMD shows the promising emotion category (binary or 7-class) separability in (c).

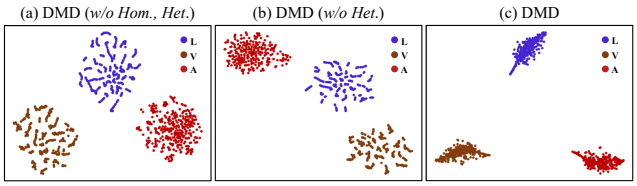


Figure 4. Visualization of the decoupled heterogeneous features on MOSEI. DMD shows the best modality separability in (c).

illustrates unsatisfactory results because the heterogeneity and distribution gap across modalities hinder the learning of the modality fusion in it. As a comparison, the multimodal features into DMD are decoupled into modality-irrelevant/-exclusive spaces. For the latter space, we use multimodal transformer to bridge the distribution gap and align the high-level semantics, thereby reducing the burden of absorbing knowledge from the heterogeneous features.

4.2. Ablation study

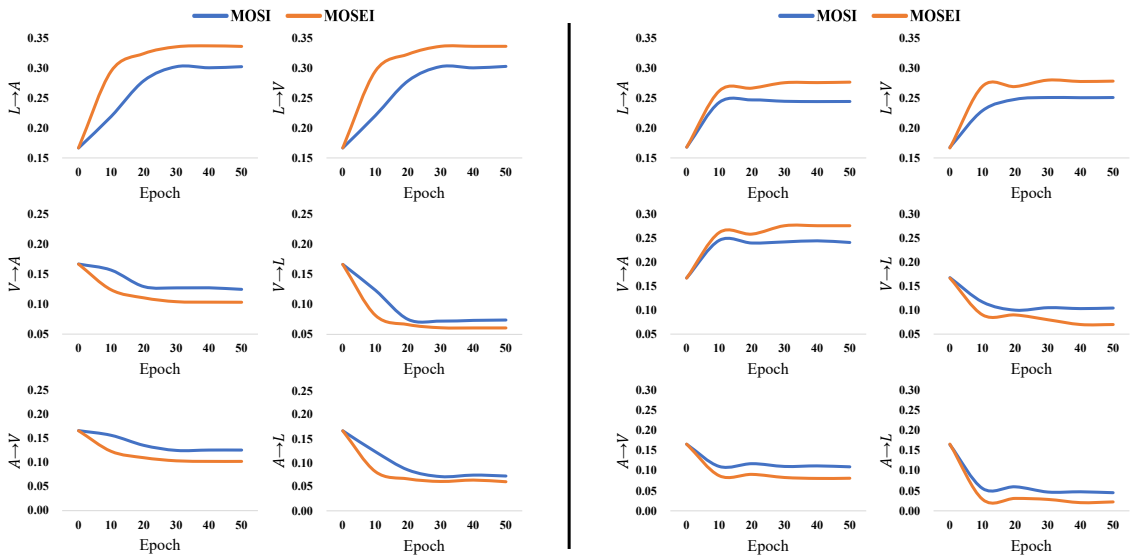
Quantitative analysis. We evaluate the effects of DMD's key components, including feature decoupling (FD), HomoGD, crossmodal attention unit (CA), HeteroGD. The results are illustrated in Tab. 3. We conclude the observations below.

Firstly, FD enhances MER performance significantly, it indicates the decoupled and refined features can reduce information redundancy and provide discriminative multimodal features. To further prove the effectiveness of FD, we conduct experiments on our baseline model with and without FD on MOSEI dataset. As shown in Tab. 4, FD brings consistent improvements for each unimodality. Meantime, the performance gap for the three modalities is reduced as the standard deviations of ACC_2 and F1 are both decreased. **Secondly**, combing FD with HomoGD brings further benefits. Although the homogeneous features are embedded in the same-dimension space, there still exists different discriminative capabilities for the modalities. HomoGD can improve the weak modalities through GD. To verify this, we conduct experiments with or without HomoGD on MOSEI dataset. The ACC_2 results are: 80.9% vs. 82.4% for *language*, 56.5% vs. 61.8% for *vision*, 54.4% vs. 64.1% for *audio*. However, conducting HeteroGD without the crossmodal attention units will generate degraded performance, indicating the multimodal transformer plays a key role in bridging the multimodal distribution gap. **Thirdly**, with CA units and HeteroGD, DMD obtains conspicuous improvements, suggesting the importance of the taking advantage of the modality-exclusive features for robust MER.

Besides, we compare our proposed DMD with the classical MulT [28] for further investigation. The results are shown in Tab. 5, where MulT (*w/ GD*) means we add a GD-Unit on MulT to conduct adaptive knowledge with the reinforced multimodal features. Essentially, the core difference between MulT (*w/ GD*) and DMD is that DMD incorporates feature decoupling. The quantitative comparison in Tab. 5 shows that DMD obtains consistent improvements than MulT (*w/ GD*). It suggests decoupling the multimodal features before distillation is feasible and reasonable. Furthermore, DMD achieves more significant improvements than the vanilla MulT, indicating the benefits of combing the feature decoupling and the graph distillation mechanisms.

Visualization of the decoupled features. We visualize the decoupled homogeneous and heterogeneous features of DMD, DMD (*w/o Hom., Het.*), DMD (*w/o Het.*) in Fig.3 and Fig.4 for a quantitative comparison. DMD (*w/o Hom., Het.*) denotes DMD without HomoGD and HeteroGD. Besides, DMD (*w/o Het.*) means DMD without HeteroGD. To visualize the homogeneous features, we randomly select 28 samples (four samples for each emotion category) in the testing set of the CMU-MOSEI dataset For the heterogeneous features, we randomly select 400 samples in the testing set of the CMU-MOSEI dataset. The features of the selected samples are projected into a 2D space by t-SNE.

For the homogeneous multimodal features of DMD and DMD (*w/o Het.*), the samples belonging to the same emotion category naturally cluster together due to their intermodal homogeneity. With the decoupled homogeneous fea-



(a) Graph edges in HomoGD (b) Graph edges in HeteroGD

Figure 5. Illustration of the graph edges in HomoGD and HeteroGD. In (a), $L \rightarrow A$ and $L \rightarrow V$ are dominated because the homogeneous language features contribute most and the other modalities perform poorly. In (b), $L \rightarrow A$, $L \rightarrow V$, and $V \rightarrow A$ are dominated. $V \rightarrow A$ emerges because the *visual* modality enhanced its feature discriminability via the multimodal transformer mechanism in HeteroGD.

tures but without graph distillation mechanism in DMD (*w/o Hom., Het.*), the samples merely show basic separability for the binary *non-negative* and *negative* categories. However, the samples are not distinguishable under the 7-class setting, indicating the features are not so discriminative than that of DMD or DMD (*w/o Het.*). The comparison between DMD (*w/o Hom., Het.*) and DMD, and the comparison between DMD (*w/o Hom., Het.*) and DMD (*w/o Het.*) verifies the effectiveness of the graph distillation on the homogeneous multimodal features.

In the heterogeneous space, due to its inter-modal heterogeneity, the features of different samples should cluster by modalities. As shown in Fig 4, DMD shows the best feature separability, indicating the complementarity between modalities is mostly enhanced. DMD (*w/o Hom., Het.*) and DMD (*w/o Het.*) show less feature separability than DMD, indicating the importance of the graph distillation on the heterogeneous multimodal features.

Visualization of graph edges in the GD-Units. We show the dynamic edges in each GD-Unit in Fig. 5 for analysis. Each graph edge corresponds to the strength of a directed distillation. We conclude the observations as: (1) The distillation in HomoGD are mainly dominated by $L \rightarrow A$ and $L \rightarrow V$. This is because the decoupled homogeneous language modality still plays the most critical role and outperforms visual or acoustic modality with significant advantages. For binary MER on the CMU-MOSEI dataset, *language*, *visual*, *acoustic* modality respectively obtains 80.9%, 56.5%, 54.4% accuracy using the decoupled homogeneous features. (2) For HeteroGD, $L \rightarrow A$, $L \rightarrow V$, and $V \rightarrow A$ are dominated. An interesting phenomenon is that $V \rightarrow A$ emerges. This should be reasonable because the *visual* modality enhanced its feature discriminability via the multimodal transformer mechanism in HeteroGD. Actually, the three modalities obtain 84.5%, 83.8%, 71.0% accuracy, respectively. Conclusively, the graph edges learn meaningful patterns for adaptive crossmodal distillation.

5. Conclusion and discussion

Within this paper we have proposed a decoupled multimodal distillation method (DMD) for MER. Our method is inspired by the observation that the contribution of different modalities varies significantly. Therefore, robust MER can be achieved by distilling the reliable and generalizable knowledge across the modalities. To mitigate the modality heterogeneities, DMD decouples the modal features into modality-irrelevant/-exclusive spaces in a self-regression manner. Two GD-Units are incorporated for each decoupled features to facilitate adaptive cross-modal distillation. Quantitative and qualitative experiments consistently demonstrate the effectiveness of DMD. A limitation is that DMD does not explicitly consider the intra-modal interactions. We will explore this in future work.

Acknowledgement: This work was supported by the National Natural Science Foundation of China (62102180, 62072244), the Natural Science Foundation of Jiangsu Province (BK20210329), Shuangchuang Program of Jiangsu Province(JSSCBS20210210).

References

[1] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, 2022. 3

[2] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016. 5

[3] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep—a collaborative voice analysis repository for speech technologies. In *ICASSP*, pages 960–964. IEEE, 2014. 6

[4] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616, 2018. 3

[5] Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. Online knowledge distillation via collaborative learning. In *CVPR*, pages 11020–11029, 2020. 3

[6] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016. 2

[7] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131, 2020. 2, 3, 5, 6

[8] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *CVPR*, pages 1921–1930, 2019. 3

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS workshop*, 2015. 3

[10] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, pages 4171–4186, 2019. 5

[11] Yong Li, Jiabei Zeng, and Shiguang Shan. Learning representations for facial actions from unlabeled videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):302–317, 2020. 5

[12] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Self-supervised representation learning from videos for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer vision and pattern recognition*, pages 10924–10933, 2019. 5

[13] Tao Liang, Guosheng Lin, Lei Feng, Yan Zhang, and Fengmao Lv. Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion. In *ICCV*, pages 8148–8156, 2021. 1, 2, 5, 6

[14] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2018. 2, 6

[15] Zhentao Liu, Min Wu, Weihua Cao, Luefeng Chen, Jianping Xu, Ri Zhang, Mengtian Zhou, and Junwei Mao. A facial expression emotion recognition based human-robot interaction system. *IEEE/CAA Journal of Automatica Sinica*, 4(4):668–676, 2017. 1

[16] Zelun Luo, Jun-Ting Hsieh, Lu Jiang, Juan Carlos Niebles, and Li Fei-Fei. Graph distillation for action detection with privileged modalities. In *ECCV*, pages 166–183, 2018. 3

[17] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *CVPR*, pages 2554–2562, 2021. 1, 2, 3, 5, 6

[18] Prem Melville, Wojciech Gryc, and Richard D Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1275–1284, 2009. 1

[19] Soma Minami, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Knowledge transfer graph for deep collaborative learning. In *ACCV*, 2020. 3

[20] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *AAAI*, pages 5191–5198, 2020. 3

[21] Dang Nguyen, Sunil Gupta, Trong Nguyen, Santu Rana, Phuoc Nguyen, Truyen Tran, Ky Le, Shannon Ryan, and Svetha Venkatesh. Knowledge distillation with distribution mismatch. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 250–265. Springer, 2021. 2

[22] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, 2019. 3

[23] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 5

[24] Sintija Petrovica, Alla Anohina-Naumeca, and Hazim Kemal Ekenel. Emotion recognition in affective tutoring systems: Collection of ground-truth data. *Procedia Computer Science*, 104:437–444, 2017. 1

[25] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899, 2019. 2

[26] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *AAAI*, pages 6892–6899, 2019. 6

[27] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *NeurIPS*, 2015. 3

[28] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov.

Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2019. 1, 2, 3, 5, 6, 7

[29] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *ICLR*, 2019. 2, 3, 6

[30] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *AAAI*, pages 7216–7223, 2019. 6

[31] Yang Wu, Zijie Lin, Yanyan Zhao, Bing Qin, and Li-Nan Zhu. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4730–4738, 2021. 2

[32] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1642–1651, 2022. 2, 3, 5, 6

[33] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Empirical Methods in Natural Language Processing, EMNLP*, 2017. 2, 6

[34] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *AAAI*, 2018. 2

[35] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016. 5

[36] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2236–2246, 2018. 5, 6

[37] Bo Zhang, Xiaoming Zhang, Yun Liu, Lei Cheng, and Zhoujun Li. Matching distributions between model and data: Cross-domain knowledge distillation for unsupervised domain adaptation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5423–5433, 2021. 2

[38] Chenrui Zhang and Yuxin Peng. Better and faster: knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification. In *IJCAI*, 2018. 3

[39] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *CVPR*, 2022. 3