TDMER

A Task-Driven Method for Multimodal Emotion Recognition

Xu, Qian; Gu, Yu; Li, Chenyu; Zhang, He; Lin, Hai Xiang; Liu, Linsong

DOI
10.1109/ICASSP49660.2025.10889666

Publication date
2025

Document Version
Final published version

Published in
ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings

ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10889666

TDMER: A Task-Driven Method for Multimodal Emotion Recognition

Qian Xu*, Yu Gu*⋆, Chenyu Li*, He Zhang†, Hai-Xiang Lin‡, Linsong Liu*
*School of Artificial Intelligence, Xidian University, Xi'an, China.
†School of Journalism and Communication, Northwest University, Xi'an, China.
‡Delft Institute of Applied Mathematics, Delft University of Technology, Delft, The Netherlands.

***Abstract*—In multimodal emotion recognition, disentangled representation learning method effectively address the inherent heterogeneity among modalities. To facilitate the flexible integration of enhanced disentangled features into multimodal emotional features, we propose a task-driven multimodal emotion recognition method TDMER. Its Cross-Modal Learning module promotes adaptive cross-modal learning of features disentangled into modality-invariant and modality-specific subspaces, based on their contributions to emotional classification probabilities. The Task-Contribution Fusion mechanism then assigns controllable weights to the enhanced features according to their task objectives, generating multimodal fusion features that improve the emotion classifier's discriminative ability. The proposed TDMER approach has been evaluated on two widely-used multimodal emotion recognition benchmarks and demonstrated significant performance improvements compared with other state-of the-art methods.**

***Index Terms*—Disentangled Representation Learning, Cross-Modal Attention Learning, MultiModal Fusion.**

I. Introduction

With the surge in interest in affective computing [1]–[3], multimodal emotion recognition (MER) is a critical field in advancing emotion recognition and analysis by harnessing information from several modalities. The applications of MER are extensive and practical, including intelligent customer service [4], [5] and mental health research [6], [7].

Conventional MER methods generally fall into two categories: i) designing intricate fusion strategies [8]–[11] for multimodal features; ii) using cross-modal attention mechanisms [12]–[15] to enhance multimodal features. However, the data types of multimodal signals differ, as do their mechanisms for encoding emotional information. Such inter-modal heterogeneity significantly impacts the multimodal fusion. To alleviate information mismatch caused by it, a core strategy has been applied: project the features of each modality into modality-invariant and modality-specific subspaces [16], [17], and then conduct effective feature learning and fusion [18]–[20].

⋆Corresponding author. E-mail: guyu@xidian.edu.cn

This work was supported in part by the National Key Research and Development Program of China under Grant 2021ZD0110400 and Grant 2021ZD0110404; in part by the National Natural Science Foundation of China under Grant 62271377 and Grant 62201407; in part by the Key Research and Development Program of Shannxi Program under Grant 2021ZDLGY01-06, Grant 2022ZDLGY01-12, Grant 2023YBGY244, Grant 2023QCYLL28, Grant 2024GX-ZDCYL-02-08, and Grant 2024GX-ZDCYL-02-17

In cross-modal learning, Hazarika et al. [21] use attention mechanisms to push emotional information exchange between decoupled features. However, such approaches may not fully utilize each representation's unique strengths due to manually specified learning directions. For instance, if audio features owns more critical emotional cues but cross-modal learning is directed from other modalities, it may reduce model efficiency.

In the realm of cross-modal fusion, common techniques are tensor fusion [9], [22] and adaptive weight fusion [23], [24]. These methods synergize effective information across modalities to assist in emotion recognition. However, while some methods [25], [26]account for the specific contributions of each modality to the final task and adjust fusion dynamically, they still overlook that the task contributions of disentangled features can also serve as controllable weights.

Existing multimodal emotion recognition methods have two primary drawbacks: i)manually prescribing the direction of cross-modal learning is not well-suited to the complex and varied distribution of emotional features; ii)existing multimodal fusion weight allocation mechanisms lack attempts to optimize emotion prediction by using the emotional classification probabilities of decoupled features as dynamic weights directly.

To address these issues, we propose the Task-Driven Multimodal Emotion Recognition method TDMER. It introduces a dynamic Cross-Modal Learning (CML) module that leverages the logits of each modality's features to guide the direction and intensity of cross-modal learning. This module adaptively enhances the critical information necessary for emotion recognition within the decoupled common and private features. We design the Task Contribution-based Fusion (TCF) module, which dynamically assigns weights to the enhanced features based on their individual true classification probabilities (TCP), facilitating task-driven multimodal fusion.

II. Method

A. Model Overview

The architecture of our proposed TDMER is illustrated in Fig.1. It primarily consists of three modules: Feature Extraction and Decoupling, CML module, and TCF mechanism. Firstly, we designed a feature decoupling module to project features from each modality into respective subspaces. Secondly, we implemented the CML module for adaptive cross-modal learning of decoupled modality features. Finally, we

fuse the enhanced decoupled features using weights based on their true classification probabilities (TCP) with TCF mechanism.

B. Feature Decoupling

Initially, we handle audio, text, and video modalities with 1D convolutional networks to extract temporally informative features: $X_m \in \mathbb{R}^{(L_m \times d_m)}$ for each modality $\mathrm{m} \in \{A, T, V\}$, with $d_A = d_T = d_V$, standardizing the dimensionality across modalities.

We use a common encoder ε^{com} and three private encoders ε_m^{pri} to obtain common feature C_m and private feature P_m:

$$C_m = \varepsilon^{com}(X_m), P_m = \varepsilon_m^{pri}(X_m) \tag{1}$$

Then, we design following four key loss functions to guide these encoders to perform feature decoupling effectively.

We utilize the Center Moment Discrepancy (CMD) metric [27] to align features in the modality-invariant subspace. For random samples *X* and *Y* with distributions *p* and *q* on $[a, b]^N$, the CMD regularizer $(CMD)_K$ is the empirical estimate of the CMD:

$$\begin{aligned} CMD_K(X, Y) = &\frac{1}{|b-a|}||E(X) - E(Y)||_2 \\ &+ \sum_{k=2}^{K} \frac{1}{|b-a|^k}||C_k(X) - C_k(Y)||_2 \end{aligned} \tag{2}$$

where $E(X) = \frac{1}{|X|}\sum_{x \in X} x$ is the empirical expectation of *X* and $C_k(X) = E(x - E(X)^k)$ is its k^{th} order central moments.

The consistency loss L_{con} is computed between modalities:

$$L_{con} = \sum_{(m_1, m_2) \in \{(a,t),(a,v),(t,v)\}} CMD_K(C_{m_1}, C_{m_2}) \tag{3}$$

To boost reconstruction capabilities, We concatenate the C_m and P_m of each modality and utilize private decoders to obtain features $D_m[C_m, P_m]$. They are compared with original features to ensure that the decoupled features can be restored as much as possible. The restruction loss L_{rec} is defined as:

$$L_{rec} = \sum ||X_m - D_m(C_m, P_m)||_2 \tag{4}$$

To ensure the decoupled features contain corresponding information instead of redundant content. We utilize soft orthogonality to constrain the similarity between common and private features, preventing excessive redundancy. Formally,

$$L_{cp} = \sum \cos(C_m, P_m) \tag{5}$$

As it is essential to avoid the model's generalization of all information as unique to each modality while comparing features. Therefore, we compare the private features of two modalities to reduce the similarity and redundancy of private information between different modalities:

$$L_{pp} = \sum_{(m_1, m_2) \in \{(a,t),(a,v),(t,v)\}} \cos(P_{m_1}, P_{m_2}) \tag{6}$$

C. Cross-Modal Learning

We have designed respective cross-modal learning modules for the common and private features respectively as follows.

For the learning of common features, we introduce the weight $\omega_{p \to q}$ to represent the strength of modality *q*'s feature learning from modality *p*, and define $\psi_{p \to q}$ to represent the difference in logits between *p* and *q*. Here, $p, q \in \{T, A, V\}$.

Then, defining $F_1(h_p, \theta_1)$ and $F_1(h_q, \theta_1)$ as the outputs of modality p and q's common feature through a fully connected (FC) layer with parameters θ_1, and further passing them through a FC layer with learnable parameters θ_2:

$$\omega_{p \to q} = Softmax(F_2([[F_1(h_p, \theta_1), h_p], [F_1(h_q, \theta_1), h_q]], \theta_2)) \tag{7}$$

The loss function for cross-modal learning is as follows:

$$L_{dc} = ||\omega \odot \psi|| \tag{8}$$

where, $\odot$ means element-wise multiplication.

For the learning of private features, we employ cross-modal attention [12], [13], [28] additionally. Taking text specific modality P_t as the source and audio modality P_a as the target, the cross-modal attention is: $Q_t = P_t P_q$, $K_a = P_a P_k$, $V_a = P_a P_v$,where P_q, P_k, and P_v are learnable parameters. Individual heads [29] $P^e_{a \to t} = Soft\max(\frac{Q_t K_a^T}{\sqrt{d}})V_a$ is the enhanced features from audio to text, where d means the dimension of Q_t and K_a.

Similarly, we can also acquire the cross-modal dynamic learning loss L_{dp} based on logits.

Meanwhile, we designed an enhanced feature loss to ensure low similarity between the augmented private features and the common features within the same modality, preventing redundant learning during cross-modal learning. Formally,

$$L_{en} = \sum \cos(C_m, P^e_m) \tag{9}$$

D. Modality Fusion

It's essential to allocate different fusion weights to enhanced decoupled features obtained earlier when composing multimodal recognition features. Therefore, we apply a classifier on each feature to obtain TCP [30], [31] as their contributions in the fusion process: $TCP_i^m = (z_i^m)I_i^*, m \in \{C, P_a, P_t, P_v\}$. Here, z_i^m is the prediction probability, and I_i^* is the index of emotion label for each utterance u_i.

In contrast to adaptive weighting that relies on optimization methods like backpropagation, using task performance indicators such as TCP as direct weights provides a more intuitive approach to feature fusion. However, since true sentiment labels are unavailable during evaluation, we utilize predicted values that closely approximate TCP post-training as feature weights: $\omega_i^{C/P_a/P_t/P_v} = Sigmoid(MLP^p_{C/P_a/P_t/P_v}(C_i/P_{ai}/P_{ti}/P_{vi})$.

Using following constraint functions to obtain the weights:

$$L_p^m = -\sum_{i=1}^{n} log(z_i^m(I_i^*)) \tag{10}$$

$$L_q^m = \sum_{i=1}^{n} MSE(TCP_i^m, \omega_i^m) \tag{11}$$

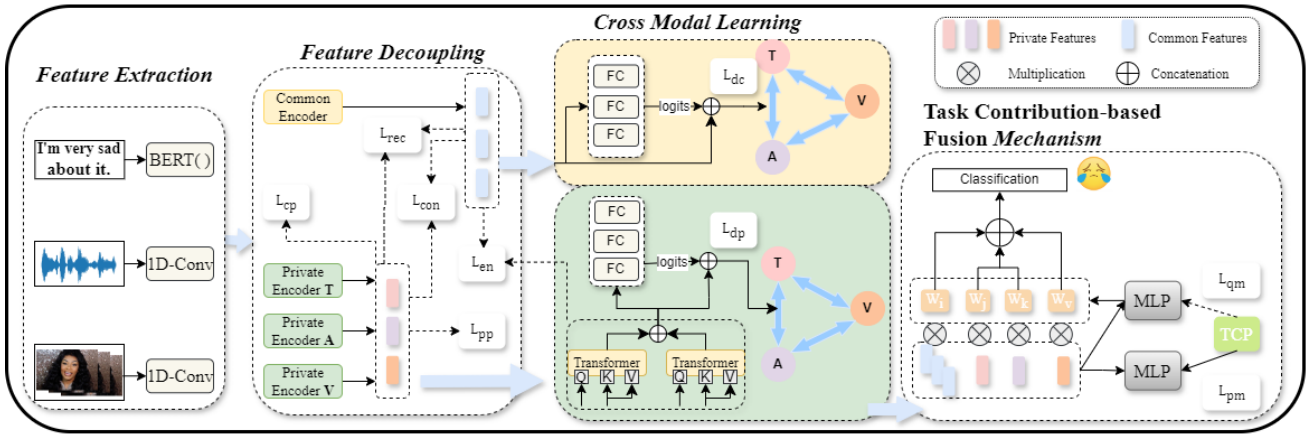

Fig. 1: The overall structure of proposed TDMER.

$$L_f = \sum_{m \in \{C, P_a, P_t, P_v\}} (L_p^m + L_q^m) \tag{12}$$

where, L_p^m represents the prediction label loss and L_p^m represents the prediction TCP loss. Then, we can get the fused multimodal features $h_i = \omega_i^C C_i + \omega_i^{p_a} p_{ai} + \omega_i^{p_t} p_{ti} + \omega_i^{p_v} p_{vi}$.

E. Objective optimization

Finally, the fused multimodal representation h_i is used for obtaining the probability of predicted emotion $y_i = Softmax(MLP(h_i))$. Then the emotion task related loss is:

$$L_{task} = L_{dc} + L_{dp} - \sum_{i=1}^{n} logy_i, I_i^* \tag{13}$$

We can define the objective by integrating above losses:

$$L = \alpha(L_{con} + L_{rec}) + \beta(L_{cp} + L_{pp}) + \gamma L_{en} + \delta L_f + L_{task} \tag{14}$$

Here, $\alpha, \beta, \gamma, \delta$ determine the importance of different constraints in the overall loss L.

III. EXPERIMENTS

A. Datasets

Our evaluations utilized the CMU-MOSI [32] and CMU-MOSEI [33] benchmark datasets for multimodal emotion recognition, both comprising audio, text, and video modalities. These datasets are annotated with sentiment scores that span the range from -3 (highly negative) to 3 (highly positive). We employed their pre-defined training, validation, and testing splits for our experiments and analysis.

B. Implementation Details

For the audio modality, we use the COVAREP [34] toolkit to extract 74-dimensional low-level features. For the visual modality, we utilize Facet [35] to encode each video frame, capturing 35 facial action units. And we use BERT [36] pre-trained model to obtain text embeddings rich in high-level semantic information. During the experiments, the batchsize was set to 16 for CMU-MOSI and 32 for CMU-MOSEI. We utilized an RTX 3090 GPU equipped with 24GB of memory and PyTorch [37] to conduct model training and evaluation.

C. Comparison with State-of-the-Art Methods

We compare the proposed TDMER model with other current state-of-the-art Multimodal Emotion Recognition(MER) methods. The TFN [9] and LMF [18] employ tensor fusion and related low-rank variants in the models. The MFM [38] is a multimodal learning model based on generative-discriminative decoupling representations. The ICCN [39] extracts two pairs of fusion features, using a correlation analysis network for emotion classification. The MulT [13] model leverages cross-modal attention learning to achieve modality interactions. The MISA [21], FDMER [23], and ConFEDE [40] fuse decoupled features with adversarial learning, cross-modal attention learning, and contrastive learning respectively.

TABLE I: Comparison on the CMU-MOSI

Methods	Acc2(%)	Acc7(%)	F1(%)	MAE
TFN [9]	80.8	34.9	80.7	0.901
LMF [18]	52.5	33.2	82.4	0.917
MFM [38]	81.7	35.4	81.6	0.877
ICCN [39]	83.0	39.0	83.0	0.860
MulT [13]	83.0	40.0	82.8	0.871
MISA [21]	83.4	42.3	83.6	0.783
FDMER [23]	84.6	44.1	84.7	0.724
ConFEDE [40]	84.2	42.3	84.1	0.742
TDMER(ours)	**86.1**	**44.6**	**86.0**	**0.712**

In the Tab. I, our TDMER model excels over other approaches on the CMU-MOSI dataset, demonstrating substantial improvements on emotion recognition. Whether for binary or multi-class classification tasks, it enhances the performance of multimodal emotion recognition through controllable learning.

In the Tab. II, our model maintains strong competitiveness, particularly in terms of MAE and Acc7. Despite the additional unimodal dataset enhances the training of ConFEDE [40], the performance gap is minimal. This underscores our model's fine-grained classification and emotion prediction capabilities, affirming the efficacy of our task-driven feature enhancement strategy. Overall, TDMER makes a significant contribution to improving performance in multimodal emotion recognition tasks.

TABLE II: Comparison on the CMU-MOSEI

Methods	Acc2(%)	Acc7(%)	F1(%)	MAE
TFN [9]	82.5	50.2	82.1	0.593
LMF [18]	82.0	48.0	82.1	0.623
MFM [38]	84.4	51.3	84.3	0.568
ICCN [39]	84.2	51.6	84.2	0.565
MulT [13]	82.5	51.8	82.3	0.580
MISA [21]	85.5	52.2	85.3	0.555
FDMER [23]	**86.1**	54.1	**85.8**	0.536
ConFEDE [40]	85.8	**54.9**	85.8	**0.522**
TDMER(ours)	85.9	54.3	85.7	0.532

TABLE V: Ablation study of main modules

Dataset	L_{en}	CML_C	CML_P	TCF	Acc2(%)	F1(%)	Acc7(%)	MAE
CMU-MOSI	✗	✓	✓	✓	85.4	85.4	44.7	0.715
	✓	✗	✓	✓	85.2	85.1	44.2	0.728
	✓	✓	✗	✓	85.0	85.0	43.6	0.736
	✓	✗	✗	✓	84.8	84.7	42.3	0.778
	✓	✓	✓	✗	85.3	85.2	43.4	0.765
	✓	✓	✓	✓	86.1	86.0	44.6	0.712
CMU-MOSEI	✗	✓	✓	✓	85.1	85.0	54.1	0.564
	✓	✗	✓	✓	84.8	84.9	54.1	0.573
	✓	✓	✗	✓	84.6	84.5	53.9	0.586
	✓	✗	✗	✓	84.1	84.0	53.5	0.589
	✓	✓	✓	✗	84.8	84.7	54.1	0.569
	✓	✓	✓	✓	85.9	85.7	54.3	0.532

Note: Here, a cross (✗) indicates the module or function has been removed, while a check mark (✓) denotes that the module has been retained.

D. Ablation Studies

To validate the influence of each module component, we conducted a thorough ablation study. As shown in Tab. III and IV, our ablation study on the four loss functions for feature decoupling, demonstrated a performance decline upon their removal, underscoring their significance for cross-modal learning and multimodal fusion. Exceptionally, eliminating the reconstruction loss paradoxically improved the Acc7. It can be inferred that the loss may restrict the model's capacity to learn finer details of the features, so the model's enhanced ability to refine multi-class categorization after its removal is logical.

TABLE III: Ablation study of feature decoupling functions on the CMU-MOSI.

Model	Acc2(%)	F1(%)	Acc7(%)	MAE
(-) L_{con}	85.4	85.2	45.2	0.736
(-) L_{rec}	85.4	85.2	**45.6**	0.734
(-) L_{cp}	85.2	84.9	32.8	0.951
(-) L_{pp}	85.2	85.2	43.3	0.751
TDMER	**86.1**	**86.0**	44.6	**0.712**

[a](-) represents removing the factors.

TABLE IV: Ablation study of feature decoupling functions on the CMU-MOSEI.

Model	Acc2(%)	F1(%)	Acc7(%)	MAE
- L_{con}	85.1	85.0	51.1	0.559
- L_{rec}	85.3	85.3	53.8	0.542
- L_{cp}	85.6	85.5	51.8	0.559
- L_{pp}	85.6	85.6	51.9	0.543
TDMER	**85.9**	**85.7**	**54.3**	**0.532**

Furthermore, we performed ablation studies on the model's core components, such as Cross-Modal Learning (CML), Task-Contribution Fusion (TCF), and essential elements within these modules, as illustrated in Tab. V.

Removing the loss function L_{en} led to a slight performance decline, indicating its benefit for private features to learn contributive private information in other modalities. The removal of the CML module in the common and private feature learning components respectively, resulted in a more pronounced decrease in model performance, with the greatest deterioration occurring when both were removed. This suggests that the CML module enhances the model's ability to acquire better emotional features through adaptive learning across decoupled modalities, leading to improved recognition outcomes. Lastly, the removal of the TCF module also prevented the TDMER model from achieving its original optimal emotion recognition results, confirming the module's positive contribution to addressing the MER task.

E. Visualization of the Decoupled Features

To showcase the effectiveness of our model's feature decoupling for projecting onto common and private subspaces, we visualized the feature distributions of the three modalities on the CMU-MOSEI dataset, as shown in Fig. 2. This visualization illustrates the transformation from the initial feature distribution to the disentangled state post-application of our model. Without our mechanism, direct attempts at decoupling fail to project the features onto different subspaces. In contrast, our approach successfully help disentangle features into common and specific attributes, highlighting its critical role in enabling effective cross-modal learning and fusion.

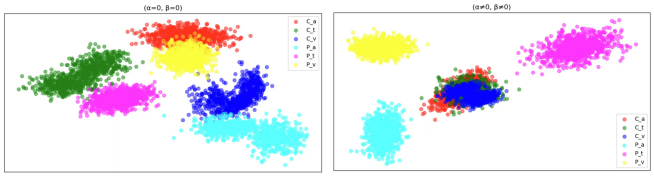

Fig. 2: Feature distribution on the CMU-MOSEI dataset before($\alpha = 0, \beta = 0$) and after($\alpha \neq 0, \beta \neq 0$) the model's feature decoupling process.

IV. Conclusion and Discussion

In this paper, we introduce TDMER, a task-driven method for multimodal emotion recognition. It decouples features into modality-invariant and modality-specific subspaces, enhancing learning and fusion through task-guided performance. TDMER employs Cross-Modal Learning module to refine decoupled features. The integration of the Task-Contribution Fusion mechanism enables dynamically merging features based on task relevance. Our evaluations reveal TDMER outperforms existing SOTA methodologies. However, considering the limitations of our model, there is still room for further improvement. As the incorporation of unimodal emotion-labeled data could be beneficial for enhancing the model's fine-grained classification capabilities, we consider to utilize relevant datasets in future work.

REFERENCES

[1] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2132–2143, 2022.

[2] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang *et al.*, "A systematic review on affective computing: Emotion models, databases, and recent advances," *Information Fusion*, vol. 83, pp. 19–52, 2022.

[3] M. M. Amin, E. Cambria, and B. W. Schuller, "Will affective computing emerge from foundation models and general artificial intelligence? a first evaluation of chatgpt," *IEEE Intelligent Systems*, vol. 38, no. 2, pp. 15–23, 2023.

[4] N. Fragopanagos and J. G. Taylor, "Emotion recognition in human–computer interaction," *Neural Networks*, vol. 18, no. 4, pp. 389–405, 2005.

[5] M. Spezialetti, G. Placidi, and S. Rossi, "Emotion recognition for human-robot interaction: Recent advances and future perspectives," *Frontiers in Robotics and AI*, vol. 7, p. 532279, 2020.

[6] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion recognition from multimodal physiological signals for emotion aware healthcare systems," *Journal of Medical and Biological Engineering*, vol. 40, pp. 149–157, 2020.

[7] J. Hu, Y. Huang, X. Hu, and Y. Xu, "The acoustically emotion-aware conversational agent with speech emotion recognition and empathetic responses," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 17–30, 2022.

[8] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE signal processing magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[9] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.

[10] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 7216–7223.

[11] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang, "Deep multimodal fusion by channel exchanging," *Advances in neural information processing systems*, vol. 33, pp. 4835–4845, 2020.

[12] D. Krishna and A. Patil, "Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks." in *Interspeech*, 2020, pp. 4243–4247.

[13] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for computational linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.

[14] T. Liang, G. Lin, L. Feng, Y. Zhang, and F. Lv, "Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8148–8156.

[15] R. G. Praveen, W. C. de Melo, N. Ullah, H. Aslam, O. Zeeshan, T. Denorme, M. Pedersoli, A. L. Koerich, S. Bacon, P. Cardinal *et al.*, "A joint cross-attention model for audio-visual fusion in dimensional emotion recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2486–2495.

[16] X. Wang, H. Chen, S. Tang, Z. Wu, and W. Zhu, "Disentangled representation learning," *arXiv preprint arXiv:2211.11695*, 2022.

[17] Y. Zhang, Y. Zhang, W. Guo, X. Cai, and X. Yuan, "Learning disentangled representation for multimodal cross-domain sentiment analysis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 10, pp. 7956–7966, 2022.

[18] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," *arXiv preprint arXiv:1806.00064*, 2018.

[19] Y. Wu, Z. Lin, Y. Zhao, B. Qin, and L.-N. Zhu, "A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis," in *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, 2021, pp. 4730–4738.

[20] J. Ye, Y. Wei, X.-C. Wen, C. Ma, Z. Huang, K. Liu, and H. Shan, "Emo-dna: Emotion decoupling and alignment learning for cross-corpus speech emotion recognition," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 5956–5965.

[21] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1122–1131.

[22] X. Yan, H. Xue, S. Jiang, and Z. Liu, "Multimodal sentiment analysis using multi-tensor fusion network with cross-modal modeling," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2000688, 2022.

[23] D. Yang, S. Huang, H. Kuang, Y. Du, and L. Zhang, "Disentangled representation learning for multimodal emotion recognition," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1642–1651.

[24] Y. Li, Y. Wang, and Z. Cui, "Decoupled multimodal distilling for emotion recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6631–6640.

[25] F. Qingyun, H. Dapeng, and W. Zhaokui, "Cross-modality fusion transformer for multispectral object detection," *arXiv preprint arXiv:2111.00273*, 2021.

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

[27] W. Zellinger, T. Grubinger, E. Lughofer, T. Natschläger, and S. Saminger-Platz, "Central moment discrepancy (cmd) for domain-invariant representation learning," *arXiv preprint arXiv:1702.08811*, 2017.

[28] V. Rajan, A. Brutti, and A. Cavallaro, "Is cross-attention preferable to self-attention for multi-modal emotion recognition?" in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4693–4697.

[29] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[30] C. Corbière, N. Thome, A. Bar-Hen, M. Cord, and P. Pérez, "Addressing failure prediction by learning model confidence," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[31] Z. Han, F. Yang, J. Huang, C. Zhang, and J. Yao, "Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20 707–20 717.

[32] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[33] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

[34] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 960–964.

[35] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–10.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[37] S. Imambi, K. B. Prakash, and G. Kanagachidambaresan, "Pytorch," *Programming with TensorFlow: solution for edge computing applications*, pp. 87–104, 2021.

[38] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," *arXiv preprint arXiv:1806.06176*, 2018.

[39] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8992–8999.

[40] J. Yang, Y. Yu, D. Niu, W. Guo, and Y. Xu, "Confede: Contrastive feature decomposition for multimodal sentiment analysis," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7617–7630.