

# MER 2023: Multi-label Learning, Modality Robustness, and Semi-Supervised Learning

Zheng Lian

Institute of Automation, Chinese Academy of Sciences (CAS)  
Beijing, China

Kang Chen

Peking University  
Beijing, China

Ke Xu

University of Chinese Academy of Sciences  
Beijing, China

Jinming Zhao

Renmin University of China  
Beijing, China

Jiangyan Yi

Institute of Automation, CAS  
Beijing, China

Guoying Zhao

University of Oulu  
Oulu, Finland

Haiyang Sun

University of Chinese Academy of Sciences  
Beijing, China

Mingyu Xu

Institute of Automation, CAS  
Beijing, China

Yu He

University of Chinese Academy of Sciences  
Beijing, China

Ye Liu

Institute of Psychology, CAS  
Beijing, China

Meng Wang

Ant Group  
Beijing, China

Björn W. Schuller

Imperial College London  
London, United Kingdom

Licai Sun

University of Chinese Academy of Sciences  
Beijing, China

Kexin Wang

Institute of Automation, CAS  
Beijing, China

Ying Li

Shandong Normal University  
Shandong, China

Bin Liu

Institute of Automation, CAS  
Beijing, China

Erik Cambria

Nanyang Technological University  
Singapore

Jianhua Tao

Tsinghua University  
Beijing, China

## ABSTRACT

The first Multimodal Emotion Recognition Challenge (MER 2023)<sup>1</sup> was successfully held at ACM Multimedia. The challenge focuses on system robustness and consists of three distinct tracks: (1) MER-MULTI, where participants are required to recognize both discrete and dimensional emotions; (2) MER-NOISE, in which noise is added to test videos for modality robustness evaluation; (3) MER-SEMI, which provides a large amount of unlabeled samples for semi-supervised learning. In this paper, we introduce the motivation behind this challenge, describe the benchmark dataset, and provide some statistics about participants. To continue using this dataset after MER 2023, please sign a new End User License Agreement<sup>2</sup> and send it to our official email address<sup>3</sup>. We believe this high-quality dataset can become a new benchmark in multimodal emotion recognition, especially for the Chinese research community.

<sup>1</sup><http://merchallenge.cn/mer2023>

<sup>2</sup><https://drive.google.com/file/d/1LOW2e6ZuyUjurVF0SNPisqSh4VzEl5LN>

<sup>3</sup>[merchallenge.contact@gmail.com](mailto:merchallenge.contact@gmail.com)



This work is licensed under a Creative Commons Attribution International 4.0 License.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada

© 2023 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0108-5/23/10.

<https://doi.org/10.1145/3581783.3612836>

## CCS CONCEPTS

- Human-centered computing → Human computer interaction (HCI).

## KEYWORDS

Multimodal Emotion Recognition Challenge (MER 2023), multi-label learning, modality robustness, semi-supervised learning

## ACM Reference Format:

Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mingyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, Ye Liu, Bin Liu, Jiangyan Yi, Meng Wang, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. 2023. MER 2023: Multi-label Learning, Modality Robustness, and Semi-Supervised Learning. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*, October 29–November 3, 2023, Ottawa, ON, Canada. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3581783.3612836>

## 1 INTRODUCTION

Multimodal emotion recognition has become an important research topic due to its wide-ranging applications in human-computer interaction. Over the past few decades, researchers have proposed various approaches [1–3]. But due to their low robustness in complex environments, existing techniques do not fully meet the demands in practice. To this end, we launch a Multimodal Emotion Recognition Challenge (MER 2023), which aims to improve system robustness

**Table 1: Statistical information for the challenge dataset (duration: hh:mm:ss).**

Partition	# of samples labeled	# of samples unlabeled	Duration
Train&Val	3373	0	03:45:47
MER-MULTI	411	0	00:28:09
MER-NOISE	412	0	00:26:23
MER-SEMI	834	73148	67:41:24

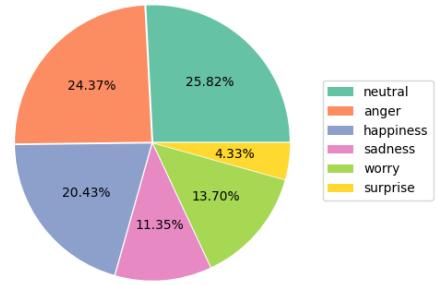
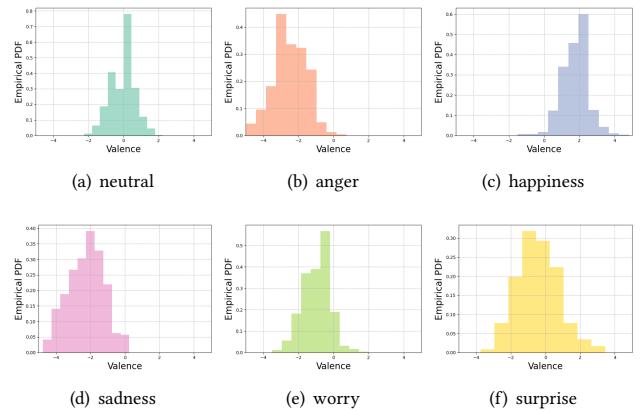
from three aspects: multi-label learning, modality robustness, and semi-supervised learning.

Annotating with both discrete and dimensional emotions is common in current datasets [4, 5]. Existing works mainly utilize multi-task learning to predict all labels simultaneously [6, 7]. However, these works ignore the correlation between discrete and dimensional emotions. For example, valence is a dimensional emotion that reflects the degree of pleasure. For negative emotions (such as *anger* and *sadness*), the valence score should be less than 0; for positive emotions (such as *happiness*), the valence score should be greater than 0. To fully exploit the multi-label correlation, we launch the MER-MULTI sub-challenge, which encourages participants to exploit the appropriate loss function [8] or model structure [9] to boost recognition performance.

Many factors may lead to modality perturbation, which increases the difficulty of emotion recognition. Recently, researchers have proposed various strategies to deal with this problem [10–12]. But due to the lack of benchmark datasets, existing works mainly rely on their own simulated missing conditions to evaluate modality robustness. To this end, we launch the MER-NOISE sub-challenge, which provides a benchmark test set focusing on more realistic modality perturbations such as background noise and blurry videos. In this sub-challenge, we encourage participants to use data augmentation [13] or other more advanced techniques [14, 15].

Meanwhile, it is difficult to collect large amounts of emotion-labeled samples due to the high annotation cost. Training with limited data harms the generalization ability of recognition systems. To address this issue, researchers have exploited various pre-trained models for video emotion recognition [16, 17]. However, task similarity impacts the performance of transfer learning [18]. Existing video-level pre-trained models mainly focus on action recognition rather than expression videos [19]. In this paper, we extract human-centered video clips from movies and TV series that contain emotional expressions. We then launch the MER-SEMI sub-challenge, encouraging participants to use semi-supervised learning [19, 20] to achieve better performance.

Therefore, MER 2023 consists of three sub-challenges: MER-MULTI, MER-NOISE, and MER-SEMI. Different from existing challenges (such as AVEC [21–29], EmotiW [30–37], and MuSE [38–40]), we mainly focus on system robustness, and provide a common platform and benchmark test sets for performance evaluation. We plan to organize a series of challenges and related workshops that bring together researchers from all over the world to discuss recent research and future directions in this field.

**Figure 1: Distribution of discrete emotions (Train&Val).****Figure 2: Empirical PDF on the valence for different discrete emotions (Train&Val).**

## 2 CHALLENGE DATASET

MER 2023 employs an extended version of CHEAVD for performance evaluation. Due to the small size of CHEAVD, we implement a fully automatic strategy to collect large amounts of unlabeled video clips; due to the low annotation consistency of CHEAVD, we adopt a stricter data selection approach and split the dataset into reliable and unreliable parts. As for reliable samples, we further divide them into three subsets: Train&Val, MER-MULTI, and MER-NOISE. As for unreliable samples, we treat them as unlabeled data and merge them with automatically-collected samples to form MER-SEMI. Statistics of each subset are shown in Table 1.

Figure 1 summarizes the distribution of discrete emotions. Despite some imbalance, our dataset still exhibits a relatively high balance compared to other mainstream benchmarks such as MELD [41] and CMU-MOSEI [5]. Figure 2 further reveals the relationship between discrete emotions and valences. Valence serves as an indicator of pleasure, and the value from small to large means the sentiment from negative to positive. From this figure, we observe that the valence distribution of different discrete labels is quite reasonable. Negative emotions (such as *anger*, *sadness*, and *worry*) predominantly exhibit valences below 0. Conversely, positive emotions (such as *happiness*) primarily exhibit valences above 0. The

**Table 2: MER-MULTI leaderboard.**

Rank	Team	Combined ( $\uparrow$ )
1	sense-dl-lab	0.7005
2	AIPL-BME-SEU	0.6860
3	USTC-qw	0.6846
4	AI4AI	0.6783
5	T_MERG	0.6765
6	SZTU-MIPS	0.6702
7	Desheng	0.6675
8	Suda_iai	0.6087
9	Emotion recognition group	0.6025
10	SUST-EiAi-Team	0.5988
11	FudanDML	0.5880
12	MCI-SCUT	0.5819
13	Beihang University	0.5713
14	ADDD	0.5673
15	Winner	0.5655
–	<b>Baseline</b>	<b>0.56</b>
16	TUA1	0.5561
17	SCUTer	0.5551
18	CiL Fighting!	0.5453
19	CCNUNLP	0.5446
20	Quaint Critters	0.5422
21	Emo.avi	0.4977
22	SDNU_AIASC	0.4639
23	Cognitist	0.1612

**Table 3: MER-NOISE leaderboard.**

Rank	Team	Combined ( $\uparrow$ )
1	sense-dl-lab	0.6846
2	AIPL-BME-SEU	0.6694
3	AI4AI	0.6371
4	LeVoice	0.6256
5	SZTU-MIPS	0.6247
6	USTC-qw	0.6162
7	Voice of Soul	0.6140
8	Desheng	0.5707
9	Beihang University	0.5462
10	SUST-EiAi-Team	0.5455
11	Triple Six	0.5444
12	FudanDML	0.5378
13	Delta	0.5339
14	USTBJDL822	0.5075
15	MCI-SCUT	0.5003
16	Trailblazers	0.4669
–	<b>Baseline</b>	<b>0.41</b>
17	ADDD	0.4055
18	CiL Fighting!	0.3863
19	Suda_iai	0.3744
20	T_MERG	0.3666
21	Quaint Critters	0.3648
22	TUA1	0.0723
23	USTC-IAT-United	-0.4608

valence associated with *neutral* centers around 0. Notably, *surprise* is a fairly complex emotion that contains multiple meanings such as sadly surprised, angrily surprised, or happily surprised. Hence, its valence ranges from negative to positive. These findings ensure the high quality of our labels and demonstrate the necessity of incorporating both discrete and dimensional annotations, as they can help us distinguish some subtle differences in emotional states.

To download the dataset, participants should fill out an End User License Agreement (EULA)<sup>4</sup>, which requires participants to use this dataset only for academic research and not to edit or upload samples to the Internet. For each track, participants can submit 20 times per day with a maximum of 200 times: MER-MULTI<sup>5</sup>, MER-NOISE<sup>6</sup>, and MER-SEMI<sup>7</sup>. At the end of the challenge, each team is required to submit a paper describing their approach. For each paper, the program committee will conduct a double-blind review of the scientific quality, novelty, and technical quality. To continue using this dataset after the challenge, please sign a new EULA<sup>8</sup> and send it to our official email address<sup>9</sup>. We will provide the test set labels to facilitate further usage. We believe this dataset can serve as a new benchmark in robust multimodal emotion recognition, especially for the Chinese research community.

<sup>4</sup><https://drive.google.com/file/d/1l0ZuPzL96W4Ow3KdF3N6D-mDhS5qJc-x>

<sup>5</sup><https://codalab.lisn.upsaclay.fr/competitions/14164>

<sup>6</sup><https://codalab.lisn.upsaclay.fr/competitions/14165>

<sup>7</sup><https://codalab.lisn.upsaclay.fr/competitions/14166>

<sup>8</sup><https://drive.google.com/file/d/1LOW2e6ZuyUjurVF0SNPisqSh4VzEl5lN>

<sup>9</sup>merchallenge.contact@gmail.com

**Table 4: MER-SEMI leaderboard.**

Rank	Team	Discrete ( $\uparrow$ )
1	sense-dl-lab	0.8911
2	SZTU-MIPS	0.8855
3	SUST-EiAi-Team	0.8853
4	Desheng	0.8841
5	AI4AI	0.8811
6	USTC-IAT-United	0.8775
7	SCUTer	0.8726
8	Voice of Soul	0.8703
9	Beihang University	0.8691
10	AIPL-BME-SEU	0.8689
–	<b>Baseline</b>	<b>0.8675</b>
11	ADDD	0.8661
12	Big Data and Intelligence Cognition	0.8537
13	TUA1	0.8507
14	T_MERG	0.8486
15	Wearing Instruments Lab	0.0661

### 3 PARTICIPANTS AND OUTCOME

This year's challenge attracts the registration of 76 teams from varying academic institutions. Due to the inherent class imbalance of discrete emotions (see Figure 1), we choose the weighted average F-score as our evaluation metric, consistent with previous works [42, 43]. For dimensional emotions, we select the widely utilized mean square errors as the evaluation metric. To further evaluate

comprehensive performance, we define a combined metric that incorporates both discrete and dimension predictions:

$$\text{metric} = \text{metric}_e - 0.25 * \text{metric}_v, \quad (1)$$

where  $\text{metric}_e$  and  $\text{metric}_v$  represent the metrics for discrete emotions and valences, respectively. In MER-MULTI and MER-NOISE, participants are required to provide predictions for both discrete and dimensional emotions. Therefore, we use the combined metric for performance evaluation. In MER-SEMI, we only evaluate discrete results on the labeled subset. Therefore, we use the weighted average F-score as the evaluation metric.

For each sub-challenge, we perform an initial attempt to explore a range of multimodal features and establish a competitive baseline system<sup>10</sup>. To ensure reproducibility, we primarily utilize open-source pre-trained models for feature extraction and a simple yet effective multi-layer perceptron for emotion recognition. In MER-MULTI and MER-NOISE, our baseline system achieves 0.56 and 0.41 on the combined metric, respectively. In MER-SEMI, we only evaluated discrete emotions and our baseline system reaches 86.40% on the weighted average F-score.

Table 2 ~ Table 4 show the leaderboards for the three sub-challenges. Excitingly, we witness that most teams exceed our baseline performance. The team named “sense-dl-lab” emerges as the winner across all three sub-challenges. Their system outperforms our baseline by 0.1405 on MER-MULTI and 0.2746 on MER-NOISE. For MER-SEMI, their system reaches 89.11% on the evaluation metric, outperforming our baseline by 2.36%.

## 4 CONCLUSIONS

This paper summarizes MER 2023, a multimodal emotion recognition challenge focused on system robustness. MER 2023 consists of three sub-challenges: (1) MER-MULTI requires participants to predict both discrete and dimensional emotions. This multi-scale labeling process can help distinguish some subtle differences in emotional states; (2) MER-NOISE simulates data corruption in real-world environments for modality robustness evaluation; (3) MER-SEMI requires participants to train more powerful classifiers using large amounts of unlabeled data. In the future, we plan to increase both labeled and unlabeled samples in our corpus. Additionally, we hope to organize a series of challenges and related workshops that bring together researchers from all over the world to discuss recent research and future directions in multimodal emotion recognition.

## 5 ACKNOWLEDGEMENTS

We would like to thank the members of data chairs Bin Liu (Institute of Automation, CAS), Ye Liu (Institute of Psychology, CAS) and Meng Wang (Ant Group). Meanwhile, we appreciate program committee for their valuable support: Shiguang Shan (Institute of Computing Technology, CAS), Jing Han (University of Cambridge), Liang Zhang (Institute of Psychology, CAS), Carlos Busso (University of Texas at Dallas), Rui Xia (Nanjing University of Science and Technology), Gualtiero Volpe (University of Genova), Yongwei Li (Institute of Automation, CAS), Giovanna Varni (Telecom Paris), Chi-Chun Lee (National Tsing Hua University), Zixing Zhang (Huazhong University), Xiaobai Li (University of Oulu), Heysem Kaya

<sup>10</sup><https://github.com/zeroQiaoba/MER2023-Baseline>

(Utrecht University), Jingming Zhao (Renmin University of China), Licai Sun (University of Chinese Academy of Sciences), Li Ya (Beijing University of Posts and Telecommunications), Mingyue Niu (Tianjin Normal University).

This work is supported by the National Natural Science Foundation of China (NSFC) (No.61831022, No.62276259, No.62201572, No.U21B2010), Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park No.Z211100004821013, Open Research Projects of Zhejiang Lab (No.2021KH0AB06), and CCF-Baidu Open Fund (No.OF2022025).

## REFERENCES

- [1] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. Recurrent neural networks for emotion recognition in video. In *Proceedings of the International Conference on Multimodal Interaction*, pages 467–474, 2015.
- [2] Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, et al. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10:99–111, 2016.
- [3] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3):55–75, 2018.
- [4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359, 2008.
- [5] 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 (Volume 1: Long Papers)*, pages 2236–2246, 2018.
- [6] Shizhe Chen, Qin Jin, Jinning Zhao, and Shuai Wang. Multimodal multi-task learning for dimensional and continuous emotion recognition. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 19–26, 2017.
- [7] Md Shad Akhtar, Dushyant Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. Multi-task learning for multi-modal emotion recognition and sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 370–379, 2019.
- [8] Huang-Cheng Chou, Chi-Chun Lee, and Carlos Busso. Exploiting co-occurrence frequency of emotions in perceptual evaluations to train a speech emotion classifier. In *Proceedings of the Interspeech*, pages 161–165, 2022.
- [9] Kexin Wang, Zheng Lian, Licai Sun, Bin Liu, Jianhua Tao, and Yin Fan. Emotional reaction analysis based on multi-label graph convolutional networks and dynamic facial expression recognition transformer. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, pages 75–80, 2022.
- [10] Jinning Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. 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 2608–2618, 2021.
- [11] Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4400–4407, 2021.
- [12] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *arXiv preprint arXiv:2208.07589*, 2022.
- [13] Devamanyu Hazarika, Yingting Li, Bo Cheng, Shuai Zhao, Roger Zimmermann, and Soujanya Poria. Analyzing modality robustness in multimodal sentiment analysis. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 685–696, 2022.
- [14] Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu. Deep partial multi-view learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(05):2402–2415, 2022.
- [15] Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [16] Zheng Lian, Jianhua Tao, Bin Liu, and Jian Huang. Unsupervised representation learning with future observation prediction for speech emotion recognition. *Proceedings of the Interspeech*, pages 3840–3844, 2019.

- [17] Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, 2022.
- [18] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.
- [19] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Proceedings of the Advances in Neural Information Processing Systems*, 2022.
- [20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15979–15988. IEEE, 2022.
- [21] Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic. Avec 2011—the first international audio/visual emotion challenge. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, pages 415–424. Springer, 2011.
- [22] Björn Schuller, Michel Valstar, Florian Eyben, Roddy Cowie, and Maja Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the International Conference on Multimodal Interaction*, pages 449–456. ACM, 2012.
- [23] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, pages 3–10, 2013.
- [24] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10, 2014.
- [25] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. Avec 2015: The 5th international audio/visual emotion challenge and workshop. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 1335–1336, 2015.
- [26] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10, 2016.
- [27] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 3–9, 2017.
- [28] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, Heysem Kaya, Maximilian Schmitt, Shahin Amiriparian, Nicholas Cummins, Denis Lalanne, Adrien Michaud, et al. Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 3–13, 2018.
- [29] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 3–12, 2019.
- [30] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, pages 509–516, 2013.
- [31] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Karan Sikka, and Tom Gedeon. Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 461–466, 2014.
- [32] Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 423–426, 2015.
- [33] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. Emotiw 2016: Video and group-level emotion recognition challenges. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 427–432, 2016.
- [34] Abhinav Dhall, Roland Goecke, Shreya Ghosh, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. From individual to group-level emotion recognition: Emotiw 5.0. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 524–528, 2017.
- [35] Abhinav Dhall, Amanjot Kaur, Roland Goecke, and Tom Gedeon. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 653–656, 2018.
- [36] Abhinav Dhall. Emotiw 2019: Automatic emotion, engagement and cohesion prediction tasks. In *Proceedings of the International Conference on Multimodal Interaction*, pages 546–550, 2019.
- [37] Abhinav Dhall, Garima Sharma, Roland Goecke, and Tom Gedeon. Emotiw 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges. In *Proceedings of the International Conference on Multimodal Interaction*, pages 784–789, 2020.
- [38] Lukas Stappen, Alice Baird, Georgios Rizos, Panagiotis Tzirakis, Xinchen Du, Felix Hafner, Lea Schumann, Adria Mallol-Ragolta, Björn W Schuller, Iulia Lefter, et al. Muse 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media: Emotional car reviews in-the-wild. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, pages 35–44, 2020.
- [39] Lukas Stappen, Alice Baird, Lukas Christ, Lea Schumann, Benjamin Sertoli, Eva-Maria Messner, Erik Cambria, Guoying Zhao, and Björn W Schuller. The muse 2021 multimodal sentiment analysis challenge: sentiment, emotion, physiological-emotion, and stress. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, pages 5–14, 2021.
- [40] Shahin Amiriparian, Lukas Christ, Andreas König, Eva-Maria Meßner, Alan Cowen, Erik Cambria, and Björn W Schuller. Muse 2022 challenge: Multimodal humour, emotional reactions, and stress. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7389–7391, 2022.
- [41] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 527–536, 2019.
- [42] Zheng Lian, Bin Liu, and Jianhua Tao. Ctret: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:985–1000, 2021.
- [43] Zheng Lian, Bin Liu, and Jianhua Tao. Decn: Dialogical emotion correction network for conversational emotion recognition. *Neurocomputing*, 454:483–495, 2021.