AVF-MAE++: Scaling Affective Video Facial Masked Autoencoders via Efficient Audio-Visual Self-Supervised Learning (Supplementary Materials)

A. Overview

In this supplementary material, we provide more details about AVF-MAE++ and present more experimental results. Specifically, we give a detailed description of the configurations of our AVF-MAE++ in Sec. B. We then present the details on building our pre-training datasets and the specifics of targeted downstream datasets in Sec. C. Afterwards, we provide more implementation details about our experiments in Sec. D. In the end, we offer more evaluation results and analysis on AVF-MAE++ in Sec. E.

B. Model Configurations

We develop three different versions of AVF-MAE++, *i.e.*, Base: AVF-MAE++ (B), Large: AVF-MAE++ (L), Huge: AVF-MAE++ (H), to extensively explore the scaling properties of audio-visual MAE for AVFA tasks. The primary differences across three model versions are about the configurations of modality-specific encoders. We display the configuration details of three models in Tab. 1 below.

C. Datasets

C.1. Unlabeled Hybrid

Our unlabeled hybrid dataset is a hybrid dataset consisting of unlabeled facial videos from CN-Celeb series [7, 21], MER2024 [52], VoxCeleb2 [14], AV-Speech [19], CelebV-HQ [111]. For all the collected facial video data, we employ the pre-processing pipeline from [21] to perform filtering and cropping, leading to further computation redundancy reduction. Note that more details about pre-processing could be found in the original paper [7]. The detailed components of the unlabeled hybrid dataset are shown in Tab. 2. Next, we specify the handing of each dataset in brief.

VoxCeleb2. VoxCeleb2 [14] includes over 1 million clips of more than 6,000 celebrities, extracted from around 150,000 interview videos on YouTube. VoxCeleb2 is divided into a development (*dev*) set and a test set. Here, we first select parts of the clips from *dev* set, and then perform pre-processing, leading to 515K processed samples.

AV-Speech. The video clips in AV-Speech [19] are collected from lectures (*e.g.*, TED Talks) and how-to videos on YouTube. Overall, this dataset contains roughly 4,700 hours of clips with approximately 150,000 distinct speakers, spanning a wide variety of people, languages and face poses. In this work, we collect parts of the overall clips and pre-process these selected clips, resulting in 371K samples.

MER2024. MER2024 is an extended version of MER2023, which consists of 115,595 video clips from Internet. After pre-processing, we obtain 85K samples for the pre-training.

CN-Celeb series. CN-Celeb series dataset is a large-scale continuous visual-speech benchmark in Mandarin Chinese, which consists of short clips collected from TV news and Internet speech shows. We randomly pick a part of data and pre-process the nonstandard clips. In the end, we take 370K clips for the pre-training stage.

CelebV-HQ. CelebV-HQ contains 35,666 video clips with the resolution of 512×512 at least, involving 15,653 identities. After data pre-processing, there exists 18K clips which could be utilized for our AVFA pre-training.

C.2. Labled Hybrid

We construct the labled hybrid datasets for the post-pre-training stage of our AVF-MAE++ on the downstream tasks to establish the progressive training pipeline by taking the union of various downstream targeted datasets.

Taking the CEA task as an example, we first align the label semantics of the downstream targeted datasets, as illustrated in Tab. 3 below. Afterwards, we remap the labels of the downstream datasets to construct annotations of the labeled hybrid dataset. Since both two datasets of the DEA task cannot apply label semantic alignment, we only construct the labeled hybrid datasets for the CEA and MER tasks. For the MER task, all of the five datasets we selected follow the popular three-emotion paradigm, so both targeted datasets and the labeled hybrid dataset adhere to the same paradigm. Notably, we do not include the MSP-IMPROV [5] dataset into the construction of CEA labeled hybrid dataset due to its unique annotation characteristics.

C.3. Targeted Fine-tuning

To verify the effectiveness and generalization ability of proposed AVF-MAE++ series models, we conduct extensive experiments on 17 datasets across three downstream AVFA tasks. The detailed information of involved datasets are provided in the following:

Configurations	AVF-MAE++ (B)	AVF-MAE++ (L)	AVF-MAE++ (H)
patch size	16	16	16
embedding dimensions (encoder)	512	640	768
number of attention heads (encoder)	8	10	12
encoder depth	10	12	15
embedding dimensions (decoder)	384	512	640
number of attention heads (decoder)	6	8	8
decoder depth	4	4	4
number of attention heads (fusion)	8	10	12
fusion depth	2	2	2
index of hierarchical skip connections [77]	[3, 6, 9]	[3, 7, 11]	[4, 9, 14]

Table 1. **The overall illustrations of configuration details about AVF-MAE++ across three different versions.** Note that the index of hierarchical skipping connections is from 0 to (encoder depth - 1).

Dataset	Size	Source
VoxCeleb2 [14]	515K	YouTube
AV-Speech [19]	371K	YouTube
MER2024 [52]	85K	Open-Media
CN-Celeb series [7, 21]	370K	Open-Media
CelebV-HQ [111]	18K	Open-Media
Unlabeled Hybrid	1.36M	Multi-Source

Table 2. **The detailed components of our unlabeled hybrid dataset.** We build this unlabeled dataset by collecting clips from multiple sources to better support AVF-MAE++ pre-training.

MAFW [56] is a multi-modal compound in-the-wild affective dataset. It consists of 10,045 clips annotated with 11 common emotions. Each video clip is also accompanied by several textual sentences to describe the subject's affective behaviors. The dataset provides an 11-class single-labeled set with 9,172 clips and a 43-class compound set with 8,996 clips. We follow the original paper to adopt a 5-fold cross-validation protocol to evaluate model performance.

DFEW [40] includes 16,372 clips extracted from movies. This dataset presents several challenging characteristics, such as extreme illumination and occlusion. We perform 5-fold cross-validation on 11,697 single-labeled clips for evaluations to align with previous works.

MER-MULTI [51] provides 3,373 training clips originating from Chinese TV series and movies. We follow the original paper to conduct 5-fold cross-validation on 3,373 clips for hyper-parameter tuning and evaluate performance on a held-out test set with 411 clips.

MER24-T&V [52] (MER2024-Train&Val) is extended by merging all the labeled samples of MER2023 Challenge [51]. Following the original paper, we fine-tune models on the *Train* set and evaluate performance on the *Val* set.

IEMOCAP [4] contains roughly 12 hours of videos from 10 subjects recorded in five sessions. Following common practice [77], we utilize 5,531 samples of five emotional categories (*i.e.*, Anger, Neutral, Happiness, Excitement, Sadness), then merge Excitement into Happiness to formulate a four-emotion form. Furthermore, we conduct 5-fold cross-validation in a session-independent manner.

CREMA-D [6] is a high-quality dataset for analysing the multi-modal patterns of acted emotions. It consists of 7,442 clips recorded by 91 actors. Since there is no official split, we follow the previous works [77, 85] to conduct 5-fold cross-validation in a *subject-independent* manner. We also conduct experiments on a subset of four emotions (*i.e.*, Happiness, Sadness, Anger, Neutral) and only report the performance of the last fold under this setup.

RAVDESS [60] consists of emotional speech and songs, which comprises 2,880 clips featuring 24 professional actors. Following [77], we only use the speech part and adopt a *subject-independent* 6-fold cross-validation protocol for evaluating performance [76, 77].

MSP-IMPROV [5] is an acted audio-visual corpus to explore emotional behaviors during conversational dyadic interactions. MSP-IMPROV [5] contains 8,438 clips recorded in six sessions from 12 actors. Following [77, 85], we only use samples of four emotions and conduct 6-fold cross-validation in a *session-independent* manner.

AVCAffe [72] is a large-scale audio-visual affect dataset simulating the remote work scenarios, which includes almost 108 hours of videos along with self-reported labels for cognitive load and affect (*i.e.*, Arousal, Valence). Since the arousal and valence scores are given on a scale of 1-4, we follow [72, 77] to formulate their predictions as a classification task. We deploy the official split (86 subjects for training and 20 subjects for test) to evaluate performance.

Werewolf-XL [102] is a database for studying spontaneous emotions during competitive group interactions of Werewolf games. It contains roughly 15 hours of audio-visual recordings. To keep consistent with previous works, we use 14,632 samples with dimensional annotations (*i.e.*, Arousal, Valence, Dominance) and conduct *subject-independent* 5-fold cross-validation for performance assessment.

CASME II [95] includes videos of 24 subjects, totaling 145

Dataset	Emotional Labels
MAFW [56]	Anger, Disgust, Fear, Happiness, Neutral, Sadness Surprise, Contempt, Anxiety, Helplessness, Disappointment
DFEW [40]	Happy, Sad, Neutral, Angry, Surprise, Disgust, Fear
MER-MULTI [51] & MER24-T&V [52]	Worried, Happy, Neutral, Angry, Surprise, Sad
IEMOCAP [4]	Anger, Happy, Neutral, Sad
CREMA-D [6]	Anger, Disgust, Fear, Happy, Neutral, Sadness
RAVDESS [60]	Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, Surprised
Labeled Hybrid	Anger, Disgust, Fear, Happy, Neutral, Sadness, Surprise Worried, Clam, Contempt, Anxiety, Helplessness, Disappointment

Table 3. **The emotional labels of the built labeled hybrid and related downstream targeted datasets.** Overall, there are 13 emotional labels in the labeled hybrid dataset for the CEA task.

Configurations	Value
video encoder mask type	tube
video encoder mask ratio	0.9
audio encoder mask type	random
audio encoder mask ratio	0.8125
video input size	$3 \times 16 \times 160 \times 160$
audio input size	$1 \times 256 \times 128$
video decoder mask type	running cell
video decoder mask ratio	0.5
audio decoder mask type	random
audio decoder mask ratio	0.5
optimizer	AdamW [61]
base learning rate	1.5*e*-4
weight decay	0.05
target normalizations	Yes
loss weight factor	0.0025
Mel filterbank sequence length	128
audio augmentation	Yes
video augmentation	MultiScaleCrop
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.95$
base batch size	164
contrastive temperature	0.07
video region size	(2, 5, 10)
audio region size	(4, 4)
repeated augmentation [37]	No
learning rate schedule	cosine decay
frame difference optimization	Yes
warmup epoch	20
epoch	200
frame	16
sampling rate	4
audio sampling rate	16000
clip grading	None (B & L), 0.7 (H)

Table 4. **The pre-training settings of AVF-MAE++.** We only show the details of AVF-MAE++ (B) here for the example.

samples. All samples are captured using lab cameras, with the frame rate of 200 FPS. After merging into three emotional categories, the number of Negative, Positive, and Surprise are 88, 32, and 25, respectively.

SAMM [17] provides various facial expression data, encompassing action unit coding as well as indices for micro-expression onset, offset, and apex. All of the videos have a resolution of 2040 $\times$ 1088 pixels with a frame rate of 200 FPS. After grouping the videos into three emotions, SAMM has 92 Negative, 26 Positive, and 15 Surprise samples.

SMIC [50] consists of video data from 16 subjects, totaling 164 samples. All the samples are recorded using a lab camera with the frame rate of 100 FPS. The original frame size of each sample is 640 $\times$ 480 pixels. The sample number of Negative, Positive, and Surprise emotions are 70, 51, and 43, respectively.

CAS(ME)^3 [49] is distinguished by its inclusion of multi-source information. In this work, we only select the *Part A* to evaluate model performance, which comprises data from 100 subjects, totaling 943 samples. The samples are captured using a lab camera, and have an original resolution of 1280 $\times$ 720 pixels. The overall number of Negative, Positive, and Surprise emotional samples are 508, 64, and 201.

MMEW [3] includes both macro- and micro-expressions. It consists of 300 MEs and 900 macro-expression samples with a large resolution (*i.e.*, 1920 $\times$ 1080) at 90 FPS. Following previous works [65, 110], we merge the seven emotions into three emotions, then evaluate perfromance.

D. Implementation Details

We pre-train three versions of AVF-MAE++ on the built unlabeled hybrid dataset using a machine with 8 $\times$ NVIDIA RTX A6000 GPUs. Besides the methods we have proposed for computational efficiency improvements in the main paper, we also adapt mix-precision training at the engineering level to speed up pre-training. Following [77, 81, 91], we train the encoder with FP-16 mixed precision and the decoder with FP-32 precision to avoid the potential precision overflow risk during model pre-training. The repeated augmentation for video data is not adapted for pre-training. The learning rate is linearly scaled according to the total batch

Method	Modality	MAFW (11-class)		CREMA-D (6-class)		Werewolf-XL
		UAR	WAR	UAR	WAR	Average
AVF-MAE++ (H)	Audio	27.15	38.78	73.51	73.02	33.22
AVF-MAE++ (H)	Video	42.24	55.61	79.23	79.97	45.38

Table 5. **Results of uni-modal AVF-MAE++ across three representive CEA and DEA datasets.** Note that we average the mertics for the three dimensions of Werewolf-XL [102] dataset.

size (*i.e.* lr = base_lr × batch_size / 256). The detailed pre-training settings are illustrated in Tab. 4 above.

In the supervised post-pre-training stage, we fine-tune the pre-trained encoder on the lableled hybrid dataset for different downstream AVFA tasks. To better maintain the pre-training effects, we slightly increase the drop path rate and adapt the repeated augmentation. Afterwards, we conduct the specific fine-tuning to output the targeted models on categorical emotion analysis (CEA), dimensional emotion analysis (DEA), and micro-expression recognition (MER) three tasks. Specifically, we employ the almost same pipeline for three downstream tasks, except that we deploy the MSE Loss and add activation functions before the model head for the DEA task. The detailed information about the post pre-training and targeted fine-tuning settings of AVF-MAE++ can be found in Tab. 6. For simplicity, we omit the parameter configurations that are consistent with the pre-training phase in Tab. 4.

E. More Evaluations

E.1. Performance Comparisons

To more comprehensively verify the effectiveness of AVF-MAE++, we first present the performance of 11 single-labeled emotions and the overall metrics on MAFW (11-class) dataset, as illustrated in Tab. 11. As can be seen, our method exhibits outstanding performance across most emotions, indicating its powerful learning capacity and the effectiveness of scaling audio-visual MAE.

Subsequently, we present more comparative results of AVF-MAE++ and other state-of-the art AVFA methods on MAFW (43-class) [56], MSP-IMPROV [5], CRMEA-D (4-class) [6], DFEW [40], AVCAffe [72], IEMOCAP [4], CRMEA-D (6-class) [6], RAVDESS [60], MER-MULTI [51], MER24-T&V [52], and Werewolf-XL [102] datasets, as displayed in the tables below. We draw the conclusion that our AVF-MAE++ can further improve the recognition results across multiple CEA and DEA datasets since it has more discriminative learning capability.

Finally, we present more comparison results of our AVF-MAE++ and state-of-the-art MER methods on five representative datasets, as shown in Tab. 19 below. The extensive experimental results show that our model exhibits competitive results, demonstarting the generalization capability of the learned affective representations by AVF-MAE++.

Configurations	Post Pre-training	Targeted Fine-tuning
optimizer	AdamW [61]	
base learning rate	1*e*-3	5*e*-4
weight decay	0.05	
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$	
inference segment	2	
inference crop	2	
learning rate schedule	cosine decay	
warmup epoch	5	
epoch	100	
drop path	0.1	
layer decay	0.75	
base batch size	32	
tubelet size	2	
color jitter factor	0.4	
mix up [100]	0.8	
RandAug [15]	(0, 0.25)	
label smoothing [79]	0.1	
repeated augmentation [37]	2	

Table 6. **The post pre-training and targeted fine-tuning settings of AVF-MAE++.** Here, we only show the detailed settings of AVF-MAE++ (B) on the CEA downstream task.

Source Dataset → Targeted Dataset	WAR
MSP-IMPROV → CREMA-D (4-class)	94.53%
RAVDESS → IEMOCAP	70.95%

Table 7. **The results of our cross-dataset study.**

E.2. Overall ablation studies

We supplement and integrate the systematical set of overall ablation studies to extensively explore the contributions of our improvements to model performance under fair settings, as illustrated in Tab. 8 below. We can clearly figure out that each improvement component can make positive impact on our overall approach. The parameter increase is primarily due to LGI-Former, which expands the attention parameters compared to Vanilla ViT [18], but effectively reduces FLOPs, as consistently demonstrated in [58]. With our careful design, we can lead to impressive pre-training speedup, as shown in Tab. 8. Meanwhile, we can improve fine-tuning speed by 20.63% compared to baseline on MAFW.

Method	Pre-training Dataset	Pre-training Time	MAFW WAR (%)	MER24-T&V WAR (%)
HiCMAE-B (Baseline)	VoxCeleb2-dev	115.45h	56.17	70.95
HiCMAE-B + Dual Masking	VoxCeleb2-dev	81.46h (1.42×)	54.63	68.93
Dual masking + Vanilla LGI-Former	VoxCeleb2-dev	77.45h (1.49×)	55.11	69.42
Dual masking + Improved LGI-Former	VoxCeleb2-dev	79.07h (1.46×)	56.12	70.36
+ Pre-training Data Scaling	Unlabeled Hybrid	84.23h (1.37×)	56.47	70.67
+ IAV-CL Module	Unlabeled Hybrid	85.87h (1.34×)	57.02	71.40
+ PSI Strategy {*i.e.*, **AVF-MAE++ (B)**}	Unlabeled Hybrid	85.87h (1.34×)	**57.50** (+1.33)	**72.11** (+1.16)
Model **AVF-MAE++ (L)**	Unlabeled Hybrid	88.58h (1.30×)	**59.13** (+2.96)	**72.33** (+1.38)
Scaling **AVF-MAE++ (H)**	Unlabeled Hybrid	93.52h (1.23×)	**60.24** (+4.07)	**72.28** (+1.33)

Table 8. **The comprehensive ablation studies for our introduced AVF-MAE++.**

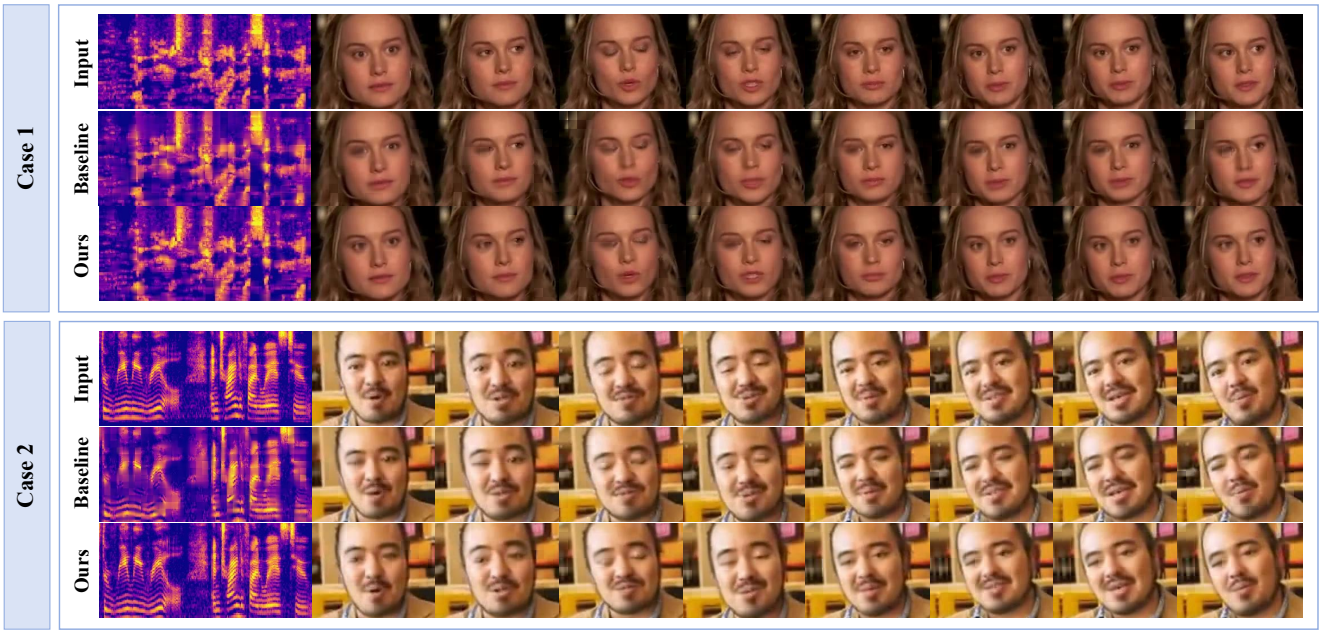

Figure 1. **The qualitative visualizations of overall audio-visual reconstructions.**

E.3. Uni-modal Comparison Results

We extra provide the uni-modal comparison results of our AVF-MAE++ (H) across three representive CEA and DEA datasets, as illustrated in Tab. 5 above. Note that please refer to Tab. 11, 10, & 17 for detailed uni-modal comparisons. As can be seen, our uni-modal models consistently exhibit competitive performance, indicating the effectiveness of our model design.

E.4. Cross-Dataset Studies

We employ the AVF-MAE++ (B) models trained on the MSP-IMPROV and RAVDESS datasets to conduct cross-dataset studies on CREMA-D (4-class) and IEMOCAP datasets, exhibiting 94.53% and 70.94% WAR, as shown in Tab. 7 above. The experimental results demonstrates that our model exhibits robust generalization and transferability of learned AVFA representations. Note that please refer to Tab. 13 & 16 below for more performance comparisons.

E.5. Qualitative Analysis

Audio-visual reconstruction visualizations. Fig. 1 shows the qualitative visualizations of overall audio-visual reconstructions. Form these outcomes, we can conclude that our method can capture crucial factors of facial emotional expressions more effectively, such as eyes and lip movements, promoting discriminative AVFA representations learning, which leads to better quality of overall reconstructions.

Confusion matrices. In Fig 2, we display the detailed confusion matrices of our proposed AVF-MAE++ (H) for five folds across MAFW (11-class) dataset. The results demonstrates that our model performs consistently and balances well across all five folds of the dataset, highlighting the robustness and generalizability of our model, thereby indicating the effectiveness of our design.

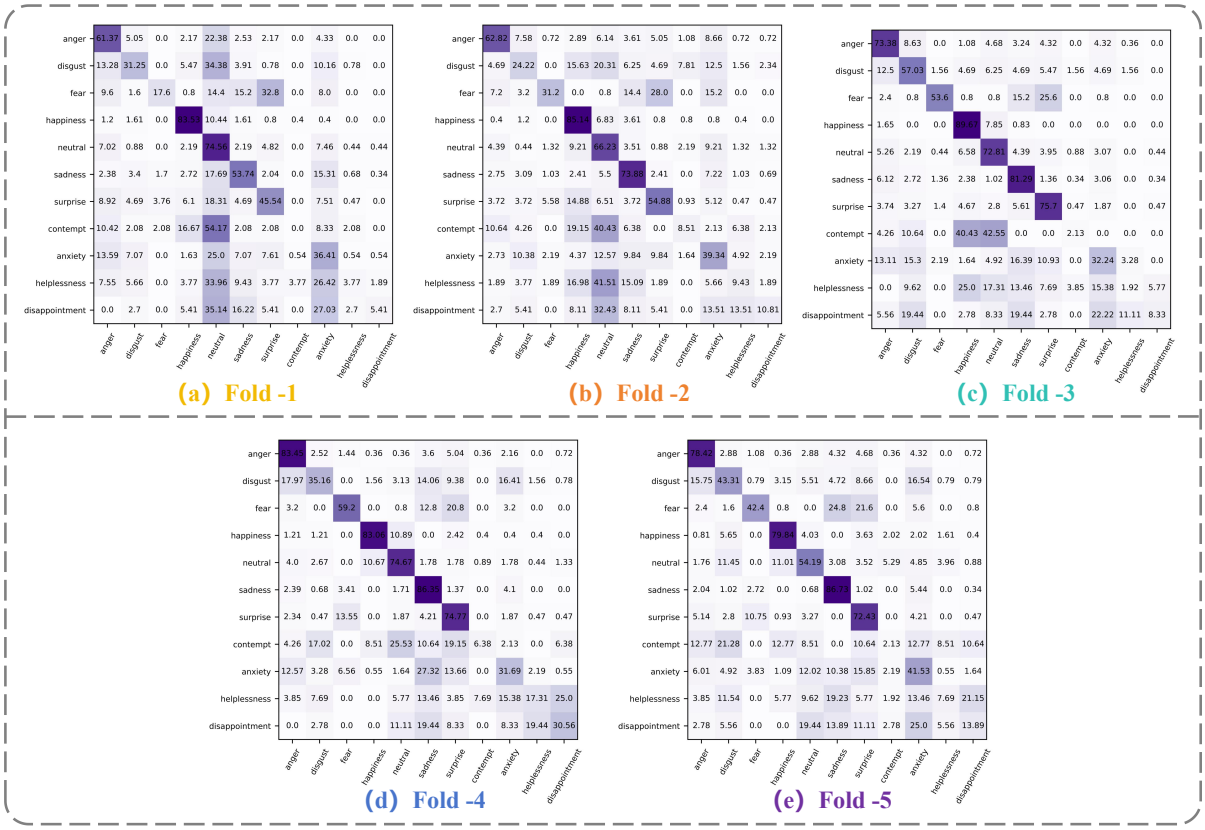

Figure 2. **The detailed confusion matrices of our AVF-MAE++ (H) for five folds across MAFW (11-class) dataset.**

Method	SSL	Modality	UAR	WAR	Macro-F1
Wav2Vec2.0 [1] (NeurIPS'20)	✓	A	5.27	20.38	–
HuBERT [38] (TASLP'21)	✓	A	5.36	20.70	–
WavLM-Plus [9] (J-STSP'22)	✓	A	5.51	21.09	–
ResNet-18 [34] (CVPR'16)	×	V	6.18	23.83	4.89
ViT [18] (ICLR'21)	×	V	8.62	31.76	7.46
C3D [82] (ICCV'15)	×	V	9.51	28.12	6.73
ResNet-18+LSTM [56] (MM'22)	×	V	6.93	26.60	5.56
ViT+LSTM [56] (MM'22)	×	V	8.72	32.24	7.59
C3D+LSTM [56] (MM'22)	×	V	7.34	28.19	5.67
T-ESFL [56] (MM'22)	×	V	9.15	34.35	7.18
Former-DFER [106] (MM'21)	×	V	10.21	32.07	–
T-MEP [104] (TCSVT'23)	×	V	9.50	31.54	–
ResNet-18+LSTM [56] (MM'22)	×	A+V	7.85	31.03	5.95
C3D+LSTM [56] (MM'22)	×	A+V	7.45	29.88	5.76
T-ESFL [56] (MM'22)	×	A+V	9.93	34.67	8.44
T-MEP* [104] (TCSVT'23)	×	A+V	11.51	34.11	–
T-MEP [104] (TCSVT'23)	×	A+V	13.22	36.58	–
HiCMAE-T [77] (IF'24)	✓	A+V	12.07	34.84	10.01
HiCMAE-S [77] (IF'24)	✓	A+V	13.47	36.29	11.53
HiCMAE-B [77] (IF'24)	✓	A+V	13.29	37.36	12.16
AVF-MAE++ (B)	✓	A+V	15.42	43.41	14.28
AVF-MAE++ (L)	✓	A+V	15.59	**43.93**	14.52
AVF-MAE++ (H)	✓	A+V	**17.25**	43.83	**15.25**
ResNet-18+MDRE [96] (SLT'18)	×	A+V+T	9.02	33.64	–
AMH [97] (ICASSP'20)	×	A+V+T	10.24	35.35	–
Rajan et al. [70] (ICASSP'22)	×	A+V+T	11.09	35.33	–
T-ESFL [56] (MM'22)	×	A+V+T	9.68	35.02	8.65
T-MEP* [104] (TCSVT'23)	×	A+V+T	13.25	37.69	–
T-MEP [104] (TCSVT'23)	×	A+V+T	15.22	39.00	–

Table 9. **The Comparative results of AVF-MAE++ with state-of-the-art methods on MAFW (43-class).** Macro-F1: macro-averaged F1-score.

Method	SSL	Modality	#Params (M)	UAR	WAR
AuxFormer [29] (ICASSP'22)	×	A	–	–	58.70
LR+eGeMAPS [20, 43]	✓	A	–	52.70	–
LR+wav2vec [1, 43]	✓	A	–	66.50	–
Wav2Vec2.0 [1] (NeurIPS'20)	✓	A	95	72.57	72.41
HuBERT [38] (TASLP'21)	✓	A	95	72.72	72.57
WavLM-Plus [9] (J-STSP'22)	✓	A	95	73.34	73.39
AuxFormer [29] (ICASSP'22)	×	V	–	–	53.10
VO-LSTM [27] (ACII'19)	×	V	–	–	66.80
Goncalves et al. [30] (TAFFC'22)	×	V	–	–	62.20
Lei et al. [45] (TAFFC'23)	×	V	–	64.68	64.76
SVFAP [78] (TAFFC'24)	✓	V	78	77.31	77.37
MAE-DFER [76] (MM'23)	✓	V	85	77.33	77.38
EF-GRU [84] (ICASSP'22)	×	A+V	–	–	57.06
LF-GRU [84] (ICASSP'22)	×	A+V	–	–	58.53
TFN [98] (EMNLP'17)	×	A+V	–	–	63.09
MATER [28] (ICIP'20)	×	A+V	–	–	67.20
MulT-Base [84] (ICASSP'22)	✓	A+V	38	–	68.87
MulT-Large [84] (ICASSP'22)	✓	A+V	89	–	70.22
AuxFormer [29] (ICASSP'22)	×	A+V	–	–	71.70
AV-LSTM [27] (ACII'19)	×	A+V	–	–	72.90
AV-Gating [27] (ACII'19)	×	A+V	–	–	74.00
Goncalves et al. [30] (TAFFC'22)	×	A+V	–	–	77.30
Ladder Networks [31] (ICASSP'23)	×	A+V	–	–	80.30
VQ-MAE-AV+ Attn. Pooling [71]	✓	A+V	30	–	78.40
VQ-MAE-AV+ Query2Emo [71]	✓	A+V	30	–	80.40
HiCMAE-T [77] (IF'24)	✓	A+V	20	83.84	83.74
HiCMAE-S [77] (IF'24)	✓	A+V	46	84.46	84.38
HiCMAE-B [77] (IF'24)	✓	A+V	81	84.91	84.89
AVF-MAE++ (B)	✓	A+V	169	85.10	85.09
AVF-MAE++ (L)	✓	A+V	303	85.69	85.60
AVF-MAE++ (H)	✓	A+V	521	**86.02**	**85.95**

Table 10. **Performance comparisons of the AVF-MAE++ with state-of-the-art methods on CREMA-D (6-class).**

Method	Venue	SSL	Modality	#Params (M)	Accuracy of Each Emotion (%)											Metrics (%)	
					AN	DI	FE	HA	NE	SA	SU	CO	AX	HL	DS	UAR	WAR
Wav2Vec2.0 [1]	NeurIPS'20	✓	A	95	59.01	9.39	26.08	31.47	32.04	46.52	9.91	1.69	12.23	3.05	6.04	21.59	29.69
HuBERT [38]	TASLP'21	✓	A	95	54.97	15.49	31.20	28.64	36.88	58.39	12.52	2.54	12.55	5.34	16.48	25.00	32.60
WavLM-Plus [9]	J-STSP'22	✓	A	95	55.62	17.21	40.48	36.65	36.53	57.44	11.12	2.12	11.35	9.54	11.54	26.33	34.07
ResNet-18 [34]	CVPR'16	×	V	11	45.02	9.25	22.51	70.69	35.94	52.25	39.04	0.00	6.67	0.00	0.00	25.58	36.65
ViT [18]	ICLR'21	×	V	–	46.03	18.18	27.49	76.89	50.70	68.19	45.13	1.27	18.93	1.53	1.65	32.36	45.04
S2D [10]	TAFFC'24	×	V	9	–	–	–	–	–	–	–	–	–	–	–	43.40	57.37
C3D [82]	ICCV'15	×	V	78	51.47	10.66	24.66	70.64	43.81	55.04	46.61	1.68	24.34	5.73	4.93	31.17	42.25
ResNet-18+LSTM [56]	MM'22	×	V	–	46.25	4.70	25.56	68.92	44.99	51.91	45.88	1.69	15.75	1.53	1.65	28.08	39.38
FE-Adapter [32]	FG'24	×	V	7	–	–	–	–	–	–	–	–	–	2.84	–	39.41	55.02
ViT+LSTM [56]	MM'22	×	V	–	42.42	14.58	35.69	76.25	54.48	68.87	41.01	0.00	24.40	0.00	1.65	32.67	45.56
C3D+LSTM [56]	MM'22	×	V	–	54.91	0.47	9.00	73.43	41.39	64.92	58.43	0.00	24.62	0.00	0.00	29.75	43.76
Former-DFER [106]	MM'21	×	V	18	58.23	11.45	31.29	75.06	43.07	63.81	46.02	0.42	26.22	2.88	2.25	32.79	45.31
T-ESFL [56]	MM'22	×	V	–	62.70	2.51	29.90	83.82	61.16	67.98	48.50	0.00	9.52	0.00	0.00	33.28	48.18
T-MEP [104]	TCSVT'23	×	V	5	52.91	17.41	28.01	80.79	49.42	58.73	49.54	0.00	26.18	2.25	3.56	33.53	47.53
DFER-CLIP [107]	BMVC'23	✓	V	153	–	–	–	–	–	–	–	–	–	–	–	39.89	52.55
SVFAP [78]	TAFFC'24	✓	V	78	64.60	25.20	35.68	82.77	57.12	70.41	58.58	8.05	32.42	8.40	9.89	41.19	54.28
EmoCLIP [23]	FG'24	✓	V	–	–	–	–	–	–	–	–	–	–	–	–	34.24	41.46
MAE-DFER [76]	MM'23	✓	V	85	67.77	25.35	34.88	77.13	58.26	71.09	57.46	**8.90**	33.08	**11.83**	12.09	41.62	54.31
UniLearn [11]	arXiv'24	✓	V	101	–	–	–	–	–	–	–	–	–	–	–	43.72	58.44
A^{3}lign-DFER [80]	arXiv'24	✓	V	–	–	–	–	–	–	–	–	–	–	–	–	42.07	53.24
ResNet-18+LSTM [56]	MM'22	×	A+V	–	54.47	11.89	7.07	82.73	54.85	55.06	39.35	0.00	15.99	0.39	0.00	29.26	42.69
C3D+LSTM [56]	MM'22	×	A+V	–	62.47	3.17	15.74	77.30	42.20	65.30	42.67	0.00	19.14	0.00	0.00	30.47	44.15
AMH [97]	ICASSP'20	×	A+V	–	51.73	18.68	28.13	79.14	52.55	52.26	46.29	0.26	29.62	1.74	2.39	32.98	48.83
T-ESFL [56]	MM'22	×	A+V	–	60.73	1.26	21.40	80.31	58.24	75.31	53.23	0.00	14.93	0.00	0.00	33.35	48.70
T-MEP* [104]	TCSVT'23	×	A+V	61	54.98	22.11	32.23	82.79	50.90	62.50	49.93	0.87	29.27	8.09	6.70	36.40	48.17
T-MEP [104]	TCSVT'23	×	A+V	61	57.04	24.85	36.09	78.96	50.83	61.85	51.28	1.29	**38.47**	6.46	1.70	37.17	51.15
MMA-DFER [13]	CVPR'24	✓	A+V	–	–	–	–	–	–	–	–	–	–	–	–	44.25	58.45
HiCMAE-T [77]	IF'24	✓	A+V	20	67.72	24.73	34.56	75.81	55.63	73.74	56.45	2.97	29.69	6.87	13.74	40.17	53.41
HiCMAE-S [77]	IF'24	✓	A+V	46	67.94	26.13	36.00	75.00	56.51	73.33	58.41	8.47	34.39	7.25	14.84	41.66	54.45
HiCMAE-B [77]	IF'24	✓	A+V	81	69.24	29.73	34.72	78.32	59.15	77.69	60.65	6.78	31.11	8.02	13.74	42.65	56.17
AVF-MAE++ (B)	–	✓	A+V	169	**76.14**	22.55	44.32	84.79	59.16	76.60	60.46	1.69	29.91	8.37	12.13	43.10	57.50
AVF-MAE++ (L)	–	✓	A+V	303	72.25	32.40	**46.56**	81.54	63.93	**78.50**	61.95	4.68	31.44	5.34	**20.33**	45.36	59.13
AVF-MAE++ (H)	–	✓	A+V	521	71.89	**38.19**	40.80	84.25	**68.49**	76.40	**64.66**	3.83	36.24	8.02	13.80	**46.05**	**60.24**
FineCLIPER [8]	MM'24	✓	T+V	20	–	–	–	–	–	–	–	–	–	–	–	45.01	56.91
ResNet18+MDRE [96]	SLT'18	×	A+T+V	–	45.59	9.35	24.30	76.31	51.10	74.87	28.82	2.08	30.99	0.00	0.00	31.22	48.33
AMH [97]	ICASSP'20	×	A+T+V	–	54.91	19.41	30.01	82.79	51.42	60.73	51.54	0.00	28.18	0.00	0.00	34.45	49.87
Rajan et al. [70]	ICASSP'22	×	A+T+V	–	56.10	9.96	41.58	84.13	60.39	63.95	44.59	0.00	24.26	2.69	1.76	35.40	48.78
T-ESFL [56]	MM'22	×	A+T+V	–	61.89	1.10	7.69	**85.90**	–	71.87	62.17	0.00	36.00	0.00	0.00	31.00	50.29
T-MEP* [104]	TCSVT'23	×	A+T+V	111	53.03	19.32	40.65	79.94	55.89	74.17	53.48	2.15	26.61	1.15	5.10	37.41	50.96
T-MEP [104]	TCSVT'23	×	A+T+V	111	56.95	18.19	42.89	81.62	60.14	71.60	58.22	3.21	30.53	2.27	7.51	39.37	52.85

Table 11. **Performance comparisons of AVF-MAE++ with state-of-the-art CEA methods on MAFW (11-class).** AN: Anger. DI: Disgust. FE: Fear. HA: Happiness. NE: Neutral. SA: Sadness. SU: Surprise. CO: Contempt. AX: Anxiety. HL: Helplessness. DS: Disappointment. UAR: Unweighted Average Recall. WAR: Weighted Average Recall. *: The pre-trained models is not deployed for initialization. –: Unaccessible Results. We highlight the best performance in **bold** and underline the second performance.

Method	SSL	Modality	#Params (M)	UAR	WAR
AuxFormer [29] (ICASSP'22)	×	A+V	–	62.97	70.28
Tran et al. [84] (ICASSP'22)	✓	A+V	–	59.41	65.29
AV-HuBERT [74] (ICLR'22)	✓	A+V	103	–	65.27
FAV-HuBERT [85] (MM'23)	✓	A+V	103	61.05	68.35
TAPT-HuBERT [85] (MM'23)	✓	A+V	103	63.95	70.46
CTAPT-HuBERT [85] (MM'23)	✓	A+V	103	60.83	68.02
AW-HuBERT [85] (MM'23)	✓	A+V	103	65.72	71.80
HiCMAE-T [77] (IF'24)	✓	A+V	20	63.16	72.78
HiCMAE-S [77] (IF'24)	✓	A+V	46	63.90	74.35
HiCMAE-B [77] (IF'24)	✓	A+V	81	65.78	74.95
AVF-MAE++ (B)	✓	A+V	169	68.90	75.07
AVF-MAE++ (L)	✓	A+V	303	68.95	75.59
AVF-MAE++ (H)	✓	A+V	521	**70.05**	**76.07**

Table 12. **The performance comparisons of our AVF-MAE++ with state-of-the-art methods on MSP-IMPROV.** #Params (M): Model parameters in million magnitude.

Method	SSL	Modality	#Params (M)	UAR	WAR
AuxFormer [29] (ICASSP'22)	×	A+V	–	91.10	91.62
Tran et al. [84] (ICASSP'22)	✓	A+V	–	83.29	83.46
AV-HuBERT [74] (ICLR'22)	✓	A+V	103	–	85.47
FAV-HuBERT [85] (MM'23)	✓	A+V	103	87.34	87.61
TAPT-HuBERT [85] (MM'23)	✓	A+V	103	92.78	92.84
CTAPT-HuBERT [85] (MM'23)	✓	A+V	103	90.52	90.39
AW-HuBERT [85] (MM'23)	✓	A+V	103	93.65	93.65
HiCMAE-T [77] (IF'24)	✓	A+V	20	92.47	92.67
HiCMAE-S [77] (IF'24)	✓	A+V	46	93.34	93.48
HiCMAE-B [77] (IF'24)	✓	A+V	81	94.00	94.13
AVF-MAE++ (B)	✓	A+V	169	93.81	94.04
AVF-MAE++ (L)	✓	A+V	303	93.05	93.16
AVF-MAE++ (H)	✓	A+V	521	**94.82**	**94.92**

Table 13. **Performance comparisons of the AVF-MAE++ with state-of-the-art methods on CREMA-D (4-class).**

Method	SSL	Modality	#Params (M)	UAR	WAR
Wav2Vec2.0 [1] (NeurIPS'20)	✓	A	95	36.15	43.05
HuBERT [38] (TASLP'21)	✓	A	95	35.98	43.24
WavLM-Plus [9] (J-STSP'22)	✓	A	95	37.78	44.64
C3D [82] (ICCV'15)	×	V	78	42.74	53.54
R(2+1)D-18 [83] (CVPR'18)	×	V	33	42.79	53.22
3D ResNet-18 [33] (CVPR'18)	×	V	33	46.52	58.27
EC-STFL [40] (MM'20)	×	V	–	45.35	56.51
ResNet-18+LSTM [106] (MM'21)	×	V	–	51.32	63.85
ResNet-18+GRU [106] (MM'21)	×	V	–	51.68	64.02
Former-DFER [106] (MM'21)	×	V	18	53.69	65.70
CEFLNet [57] (IS'22)	×	V	13	51.14	65.35
EST [58] (PR'23)	×	V	43	53.43	65.85
STT [62] (arXiv'22)	×	V	–	54.58	66.65
NR-DFERNet [46] (arXiv'22)	×	V	–	54.21	68.19
DPCNet [92] (MM'22)	×	V	51	57.11	66.32
IAL [47] (AAAI'23)	×	V	19	55.71	69.24
M3DFEL [90] (CVPR'23)	×	V	–	56.10	69.25
T-MEP [104] (TCSVT'23)	×	V	5	54.14	65.22
Video Swin-T [59] (CVPR'22)	×	V	88	59.38	71.90
UniLearn [11] (arXiv'24)	✓	V	101	66.80	76.68
CLIPER [48] (ICME'24)	✓	V	88	57.56	70.84
S2D [10] (TAFFC'24)	✓	V	9	65.45	74.81
DFER-CLIP [107] (BMVC'23)	✓	V	153	59.61	71.25
SVFAP [78] (TAFFC'24)	✓	V	78	62.83	74.27
EmoCLIP [23] (FG'24)	✓	V	–	58.04	62.12
MAE-DFER [76] (MM'23)	✓	V	85	63.41	74.43
VideoMAE [81] (NeurIPS'22)	✓	V	86	58.32	70.94
MoCo [35] (CVPR'20)	✓	V	32	53.47	67.45
A^{3}lign-DFER [80] (arXiv'24)	✓	V	–	64.09	74.20
ResNet-18+LSTM [104] (TCSVT'23)	×	A+V	–	52.41	64.32
C3D+LSTM [104] (TCSVT'23)	×	A+V	–	53.77	65.17
AMH [97] (ICASSP'20)	×	A+V	–	54.48	66.51
T-MEP* [104] (TCSVT'23)	×	A+V	61	55.06	66.30
T-MEP [104] (TCSVT'23)	×	A+V	61	57.16	68.85
HiCMAE-T [77] (IF'24)	✓	A+V	20	60.13	72.43
HiCMAE-S [77] (IF'24)	✓	A+V	46	63.05	74.33
HiCMAE-B [77] (IF'24)	✓	A+V	81	63.76	75.01
AVF-MAE++ (B)	✓	A+V	169	63.74	75.42
AVF-MAE++ (L)	✓	A+V	303	65.14	76.24
AVF-MAE++ (H)	✓	A+V	521	**66.88**	**77.45**
FineCLIPER [8] (MM'24)	✓	T+V	20	65.98	76.21

Table 14. **Performance comparisons of the AVF-MAE++ with state-of-the-art CEA methods on DFEW.**

Method	SSL	Modality	#Params (M)	Arousal	Valence
VGG-16+ MC3-18 [72] (AAAI'23)	×	A+V	47	38.90	41.70
VGG-16+ 3D ResNet-18 [72] (AAAI'23)	×	A+V	69	37.30	39.40
VGG-16+ R(2+1)D-18 [72] (AAAI'23)	×	A+V	67	40.50	39.50
ResNet-18+ MC3-18 [72] (AAAI'23)	×	A+V	44	36.00	39.20
ResNet-18+ 3D ResNet-18 [72] (AAAI'23)	×	A+V	66	35.10	39.10
ResNet-18+ R(2+1)D-18 [72] (AAAI'23)	×	A+V	64	39.50	37.70
HiCMAE-T [77] (IF'24)	✓	A+V	20	39.64	36.74
HiCMAE-S [77] (IF'24)	✓	A+V	46	42.13	42.65
HiCMAE-B [77] (IF'24)	✓	A+V	81	43.18	44.20
AVF-MAE++ (B)	✓	A+V	169	43.02	46.93
AVF-MAE++ (L)	✓	A+V	303	45.21	47.83
AVF-MAE++ (H)	✓	A+V	521	**47.25**	**49.66**

Table 15. **Performance comparisons of the AVF-MAE++ with state-of-the-art methods on AVCAffe.** The evaluation metrics for Arousal and Valence both are weighted F1-score (WA-F1).

Method	SSL	Modality	#Params (M)	UAR	WAR
FBANK [87] (ICASSP'24)	×	A	–	–	51.52
AV-HuBERT [74] (arXiv'22)	✓	A	90	–	58.54
RepLAI [64] (NeurIPS'22)	✓	A	5	–	57.53
AVBERT [44] (ICLR'21)	✓	A	10	–	60.94
MAViL [44] (ICLR'21)	✓	A	86	–	59.46
Wav2Vec2.0 [1] (NeurIPS'20)	✓	A	95	69.88	67.32
HuBERT [38] (TASLP'21)	✓	A	95	68.33	66.34
WavLM-Plus [9] (J-STSP'22)	✓	A	95	68.64	67.12
HOG [16] (CVPR'05)	×	V	–	–	35.83
AV-HuBERT [74] (arXiv'22)	✓	V	103	–	26.59
RepLAI [64] (NeurIPS'22)	✓	V	15	–	40.72
AVBERT [44] (ICLR'21)	✓	V	37	–	45.80
MAViL [44] (ICLR'21)	✓	V	87	–	43.03
AV-HuBERT [74] (ICLR'22)	✓	A+V	103	–	46.45
AVBERT [44] (ICLR'21)	✓	A+V	43	–	61.87
MAViL [44] (ICLR'21)	✓	A+V	187	–	54.94
HiCMAE-T [77] (IF'24)	✓	A+V	20	66.85	66.62
HiCMAE-S [77] (IF'24)	✓	A+V	46	67.46	67.49
HiCMAE-B [77] (IF'24)	✓	A+V	81	68.21	68.36
AVF-MAE++ (B)	✓	A+V	169	69.53	71.47
AVF-MAE++ (L)	✓	A+V	303	69.86	71.65
AVF-MAE++ (H)	✓	A+V	521	**72.71**	**73.83**

Table 16. **Performance comparisons of the AVF-MAE++ with state-of-the-art CEA methods on IEMOCAP.**

Method	SSL	Modality	Arousal	Valence	Dominance
eGeMAPS [20] (TAFFC'15)	×	A	23.45	8.08	31.15
VGGish [36] (ICASSP'17)	×	A	22.88	5.69	29.59
HiCMAE-T [77] (IF'24)	✓	A	26.54	12.94	37.88
HiCMAE-S [77] (IF'24)	✓	A	28.40	15.46	37.83
HiCMAE-B [77] (IF'24)	✓	A	30.04	17.63	36.60
HOG [16] (CVPR'05)	×	V	20.82	52.54	24.76
VGGFace [67] (BMVC'15)	×	V	7.24	62.96	14.30
SVFAP [78] (TAFFC'24)	✓	V	23.51	67.11	34.61
HiCMAE-T [77] (IF'24)	✓	V	22.45	66.55	33.57
HiCMAE-S [77] (IF'24)	✓	V	23.11	67.05	34.00
HiCMAE-B [77] (IF'24)	✓	V	24.04	67.03	34.91
Zhang et al. [102] (TAFFC'23)	×	A+V	16.41	63.14	35.40
HiCMAE-T [77] (IF'24)	✓	A+V	30.47	68.50	42.37
HiCMAE-S [77] (IF'24)	✓	A+V	31.08	68.92	41.38
HiCMAE-B [77] (IF'24)	✓	A+V	33.74	69.23	40.66
AVF-MAE++ (B)	✓	A+V	44.33	71.22	**52.59**
AVF-MAE++ (L)	✓	A+V	43.54	72.09	52.07
AVF-MAE++ (H)	✓	A+V	**44.99**	**72.19**	52.35

Table 17. **Performance comparisons of the AVF-MAE++ with state-of-the-art methods on Werewolf-XL.**

Method	SSL	Modality	#Params (M)	WAR	WA-F1
Whisper [69] (ICML'23)	×	A	1550	63.27	63.23
eGeMAPS [20] (TAFFC'15)	×	A	–	42.88	39.68
VGGish [36] (ICASSP'17)	×	A	–	50.20	48.60
emotion2vec [63] (ACL'24)	✓	A	94	56.48	56.08
Wav2Vec2.0 [1] (NeurIPS'20)	✓	A	95	65.83	65.50
HuBERT [38] (TASLP'21)	✓	A	95	69.43	69.26
EmoNet [42] (JMUI'16)	×	V	–	53.18	51.76
SENet-FER2013 [39] (CVPR'18)	×	V	28	58.79	57.67
ResNet-FER2013 [34] (CVPR'16)	×	V	26	59.66	58.73
MANet-RAFDB [108] (TIP'21)	×	V	51	61.10	59.91
CLIP-base [68] (ICML'21)	✓	V	–	62.56	61.74
CLIP-large [68] (ICML'21)	✓	V	–	67.17	66.66
EVA-02 [22] (IVC'24)	✓	V	86	62.28	61.41
DINOv2 [66] (arXiv'23)	✓	V	–	59.57	58.44
VideoMAE [81] (NeurIPS'22)	✓	V	86	64.93	64.50
HiCMAE [77] (IF'24)	✓	A+V	81	70.95	70.18
AVF-MAE++ (B)	✓	A+V	169	72.11	71.24
AVF-MAE++ (L)	✓	A+V	303	**72.33**	71.64
AVF-MAE++ (H)	✓	A+V	521	72.28	**71.75**

Table 18. **Performance comparisons of the AVF-MAE++ with state-of-the-art methods on MER24-T&V.**

Method	Venue	SSL	SAMM		CASME II		SMIC		CAS(ME)^3		MMEW	
			UAR	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR	UF1
Traditional methods												
LBP-TOP [105]	T-PAMI'07	×	41.02	39.54	74.29	70.26	52.80	20.00	21.39	21.78	63.61	64.23
Bi-WOOF [53]	IMAGE'18	×	51.39	52.11	80.26	78.05	58.29	57.27	–	–	–	–
Deep learning methods												
AlexNet [101]	IJCNN'22	×	66.42	61.04	83.12	79.94	63.73	62.01	26.34	25.70	–	–
GoogLeNet [2]	AAAI'16	×	59.92	51.24	64.14	59.89	55.11	51.23	–	–	–	–
STSTNet [54]	FG'19	×	68.10	65.88	86.86	83.82	70.13	68.01	37.92	37.95	82.53	80.37
VGG16 [73]	Frontiers in Neuroscience'19	×	47.93	48.70	82.02	81.66	59.64	58.00	–	–	–	–
CapsuleNet [88]	FG'19	×	59.89	62.09	70.18	70.68	58.77	58.20	–	–	68.34	67.62
RCN-A [94]	TIP'20	×	67.20	76.01	81.20	85.12	64.40	63.26	38.93	39.28	–	–
EMR [55]	FG'19	×	71.52	77.54	82.09	82.93	75.30	74.61	36.56	36.13	82.66	81.49
OFF-ApexNet [26]	IMAGE'19	×	53.92	54.09	86.81	87.64	66.95	68.17	–	–	–	–
μ-BERT [65]	CVPR'23	×	**84.75**	–	89.14	90.34	**83.84**	**85.50**	61.25	56.04	–	–
FeatRef [110]	PR'22	×	71.55	73.72	88.73	89.15	70.83	70.11	34.13	34.93	–	82.11
Dual-Inception [109]	FG'19	×	56.63	58.68	85.60	86.21	67.26	66.45	–	–	–	–
SLSTT-LSTM [103]	TAFFC'22	×	64.30	71.50	88.50	90.10	72.00	74.00	–	–	–	–
HTNet [93]	Neurocomputing'24	×	81.24	81.31	95.16	**95.32**	79.05	80.49	54.15	57.67	–	**84.33**
HiCMAE [77]	Information Fusion'24	✓	–	–	90.21	92.03	81.03	80.33	–	–	–	–
AVF-MAE++ (B)	–	✓	80.43	81.58	96.67	93.58	80.56	83.23	61.06	63.18	83.96	83.76
AVF-MAE++ (L)	–	✓	81.55	**82.53**	**96.72**	94.03	81.67	83.79	**66.02**	**67.88**	85.81	83.41
AVF-MAE++ (H)	–	✓	81.01	81.62	96.54	94.11	81.64	83.55	65.88	65.34	**86.23**	**84.33**

Table 19. **The comparative results of recently state-of-the-art MER methods with AVF-MAE++ in terms of Unweighted Average Recall (UAR) and Unweighted F1-score (UF1) on five representive MER datasets.** Note that we highlight the best performance in **bold** and underline the second performance.

Method	SSL	Modality	#Params (M)	UAR	WAR
LR+eGeMAPS [20, 43]	✓	A	–	50.30	–
LR+wav2vec [1, 43]	✓	A	–	68.80	–
Wav2Vec2.0 [1] (NeurIPS'20)	✓	A	95	73.44	74.38
HuBERT [38] (TASLP'21)	✓	A	95	74.15	74.37
WavLM-Plus [9] (J-STSP'22)	✓	A	95	75.28	75.36
VO-LSTM [27] (ACII'19)	×	V	–	–	60.50
3D ResNeXt-50 [75] (arXiv'20)	×	V	26	–	62.99
SVFAP [78] (TAFFC'24)	✓	V	78	75.15	75.01
MAE-DFER [76] (MM'23)	✓	V	85	75.91	75.56
AV-LSTM [27] (ACII'19)	×	A+V	–	–	65.80
AV-Gating [27] (ACII'19)	×	A+V	–	–	67.70
MCBP [25] (EMNLP'16)	×	A+V	51	–	71.32
MMTM [41] (CVPR'20)	×	A+V	32	–	73.12
MSAF [75] (arXiv'20)	×	A+V	26	–	74.86
ERANNs [89] (PRL'22)	×	A+V	–	–	74.80
CFN-SR [24] (arXiv'21)	×	A+V	26	–	75.76
MATER [28] (ICIP'20)	×	A+V	–	–	76.30
MulT [86] (ACL'19)	×	A+V	–	–	76.60
AVT [12] (ICPR'22)	×	A+V	–	–	79.20
VQ-MAE-AV+ Attn. Pooling [71]	✓	A+V	30	–	83.20
VQ-MAE-AV+ Query2Emo [71]	✓	A+V	30	–	84.80
HiCMAE-T [77] (IF'24)	✓	A+V	20	86.26	86.11
HiCMAE-S [77] (IF'24)	✓	A+V	46	86.85	86.67
HiCMAE-B [77] (IF'24)	✓	A+V	81	**87.96**	**87.99**
AVF-MAE++ (B)	✓	A+V	169	85.09	85.07
AVF-MAE++ (L)	✓	A+V	303	86.98	87.22
AVF-MAE++ (H)	✓	A+V	521	87.44	87.57

Table 20. **Performance comparisons of the AVF-MAE++ with state-of-the-art methods on RAVDESS.**

Method	SSL	Modality	#Params (M)	UAR	WA-F1
eGeMAPS [20] (TAFFC'15)	×	A	–	–	17.28
VGGish [36] (ICASSP'17)	×	A	–	–	40.76
Wav2Vec2.0 [1] (NeurIPS'20)	✓	A	95	51.36	51.48
HuBERT [38] (TASLP'21)	✓	A	95	50.32	52.70
WavLM-Plus [9] (J-STSP'22)	✓	A	95	53.43	54.16
HuBERT-CH [99] (ICASSP'22)	✓	A	95	–	61.16
HiCMAE-T [77] (IF'24)	✓	A	8	48.35	51.33
HiCMAE-S [77] (IF'24)	✓	A	18	51.09	54.16
HiCMAE-B [77] (IF'24)	✓	A	32	51.43	55.33
ResNet-MSCeleb [34] (CVPR'16)	×	V	26	–	40.32
ResNet-ImageNet [34] (CVPR'16)	×	V	26	–	44.91
SENet-FER2013 [39] (CVPR'18)	×	V	28	–	56.69
ResNet-FER2013 [34] (CVPR'16)	×	V	26	–	57.44
MANet-RAFDB [108] (TIP'21)	×	V	51	–	56.19
HiCMAE-T [77] (IF'24)	✓	V	8	50.52	58.37
HiCMAE-S [77] (IF'24)	✓	V	18	51.53	59.25
HiCMAE-B [77] (IF'24)	✓	V	32	52.31	59.87
ResNet-FER2013+ HuBERT-CH [51] (MM'23)	✓	A+V	121	–	69.11
MANet-RAFDB+ HuBERT-CH [51] (MM'23)	✓	A+V	146	–	70.32
HiCMAE-T [77] (IF'24)	✓	A+V	20	59.91	68.56
HiCMAE-S [77] (IF'24)	✓	A+V	46	63.18	70.22
HiCMAE-B [77] (IF'24)	✓	A+V	81	64.15	71.33
AVF-MAE++ (B)	✓	A+V	169	64.87	69.56
AVF-MAE++ (L)	✓	A+V	303	66.34	70.79
AVF-MAE++ (H)	✓	A+V	521	**68.20**	**72.26**

Table 21. **Performance comparisons of the AVF-MAE++ with state-of-the-art methods on MER-MULTI.**

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 6, 7, 8, 9

[2] Pedro Ballester and Ricardo Araujo. On the performance of googlenet and alexnet applied to sketches. In *Proceedings of the AAAI conference on artificial intelligence*, 2016. 9

[3] Xianye Ben, Yi Ren, Junping Zhang, Su-Jing Wang, Kidiyo Kpalma, Weixiao Meng, and Yong-Jin Liu. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5826–5846, 2021. 3

[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. 2, 3, 4

[5] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80, 2016. 1, 2, 4

[6] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014. 2, 3, 4

[7] Chen Chen, Dong Wang, and Thomas Fang Zheng. Cn-cvs: A mandarin audio-visual dataset for large vocabulary continuous visual to speech synthesis. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1, 2

[8] Haodong Chen, Haojian Huang, Junhao Dong, Mingzhe Zheng, and Dian Shao. Finecliper: Multi-modal fine-grained clip for dynamic facial expression recognition with adapters. *arXiv preprint arXiv:2407.02157*, 2024. 7, 8

[9] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16 (6):1505–1518, 2022. 6, 7, 8, 9

[10] Yin Chen, Jia Li, Shiguang Shan, Meng Wang, and Richang Hong. From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos. *IEEE Transactions on Affective Computing*, 2024. 7, 8

[11] Yin Chen, Jia Li, Yu Zhang, Zhenzhen Hu, Shiguang Shan, Meng Wang, and Richang Hong. Unilearn: Enhancing dynamic facial expression recognition through unified pre-training and fine-tuning on images and videos. *arXiv preprint arXiv:2409.06154*, 2024. 7, 8

[12] Kateryna Chumachenko, Alexandros Iosifidis, and Moncef Gabbouj. Self-attention fusion for audiovisual emotion recognition with incomplete data. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2822–2828. IEEE, 2022. 9

[13] Kateryna Chumachenko, Alexandros Iosifidis, and Moncef Gabbouj. Mma-dfer: Multimodal adaptation of unimodal models for dynamic facial expression recognition in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4673–4682, 2024. 7

[14] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 1, 2

[15] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 4

[16] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, pages 886–893. Ieee, 2005. 8

[17] Adrian K Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap. Samm: A spontaneous micro-facial movement dataset. *IEEE transactions on affective computing*, 9(1):116–129, 2016. 3

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 4, 6, 7

[19] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018. 1, 2

[20] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015. 6, 8, 9

[21] Yue Fan, JW Kang, LT Li, KC Li, HL Chen, ST Cheng, PY Zhang, ZY Zhou, YQ Cai, and Dong Wang. Cn-celeb: a challenging chinese speaker recognition dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7604–7608. IEEE, 2020. 1, 2

[22] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149: 105171, 2024. 8

[23] Niki Maria Foteinopoulou and Ioannis Patras. Emoclip: A vision-language method for zero-shot video facial expres-

sion recognition. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10. IEEE, 2024. 7, 8

[24] Ziwang Fu, Feng Liu, Hanyang Wang, Jiayin Qi, Xiangling Fu, Aimin Zhou, and Zhibin Li. A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition. *arXiv preprint arXiv:2111.02172*, 2021. 9

[25] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468, 2016. 9

[26] Yee Siang Gan, Sze-Teng Liong, Wei-Chuen Yau, Yen-Chang Huang, and Lit-Ken Tan. Off-apexnet on micro-expression recognition system. *Signal Processing: Image Communication*, 74:129–139, 2019. 9

[27] Esam Ghaleb, Mirela Popa, and Stylianos Asteriadis. Multimodal and temporal perception of audio-visual cues for emotion recognition. In *2019 8th international conference on affective computing and intelligent interaction (ACII)*, pages 552–558. IEEE, 2019. 6, 9

[28] Esam Ghaleb, Jan Niehues, and Stylianos Asteriadis. Multimodal attention-mechanism for temporal emotion recognition. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 251–255. IEEE, 2020. 6, 9

[29] Lucas Goncalves and Carlos Busso. Auxformer: Robust approach to audiovisual emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7357–7361. IEEE, 2022. 6, 7

[30] Lucas Goncalves and Carlos Busso. Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features. *IEEE Transactions on Affective Computing*, 13(4):2156–2170, 2022. 6

[31] Lucas Goncalves and Carlos Busso. Learning cross-modal audiovisual representations with ladder networks for emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 6

[32] Shreyank N Gowda, Boyan Gao, and David A Clifton. Fe-adapter: Adapting image-based emotion classifiers to videos. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2024. 7

[33] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 8

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7, 8, 9

[35] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 8

[36] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 8, 9

[37] Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8129–8138, 2020. 3, 4

[38] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021. 6, 7, 8, 9

[39] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 8, 9

[40] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2881–2889, 2020. 2, 3, 4, 8

[41] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13289–13299, 2020. 9

[42] 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. 8

[43] Aaron Keesing, Yun Sing Koh, Vithya Yogarajan, and Michael Witbrock. Emotion recognition toolkit (ertk): Standardising tools for emotion recognition research. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9693–9696, 2023. 6, 9

[44] Sangho Lee, Youngjae Yu, Gunhee Kim, Thomas Breuel, Jan Kautz, and Yale Song. Parameter efficient multimodal transformers for video representation learning. In *9th International Conference on Learning Representations, ICLR 2021*, 2021. 8

[45] Yuanyuan Lei and Houwei Cao. Audio-visual emotion recognition with preference learning based on intended and multi-modal perceived labels. *IEEE Transactions on Affective Computing*, 14(4):2954–2969, 2023. 6

[46] Hanting Li, Mingzhe Sui, Zhaoqing Zhu, et al. Nr-dfernet: Noise-robust network for dynamic facial expression recognition. *arXiv preprint arXiv:2206.04975*, 2022. 8

[47] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Intensity-aware loss for dynamic facial expression recognition in the wild. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 67–75, 2023. 8

[48] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Cliper: A unified vision-language framework for in-the-wild facial expression recognition. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 8

[49] Jingting Li, Zizhao Dong, Shaoyuan Lu, Su-Jing Wang, Wen-Jing Yan, Yinhuan Ma, Ye Liu, Changbing Huang, and Xiaolan Fu. Cas (me) 3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2782–2800, 2022. 3

[50] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen. A spontaneous micro-expression database: Inducement, collection and baseline. In *2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg)*, pages 1–6. IEEE, 2013. 3

[51] Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mngyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, et al. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9610–9614, 2023. 2, 3, 4, 9

[52] Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, et al. Mer 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. *arXiv preprint arXiv:2404.17113*, 2024. 1, 2, 3, 4

[53] Sze-Teng Liong, John See, KokSheik Wong, and Raphael C-W Phan. Less is more: Micro-expression recognition from video using apex frame. *Signal Processing: Image Communication*, 62:82–92, 2018. 9

[54] Sze-Teng Liong, Yee Siang Gan, John See, Huai-Qian Khor, and Yen-Chang Huang. Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–5. IEEE, 2019. 9

[55] Yuchi Liu, Heming Du, Liang Zheng, and Tom Gedeon. A neural micro-expression recognizer. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–4. IEEE, 2019. 9

[56] Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 24–32, 2022. 2, 3, 4, 6, 7

[57] Yuanyuan Liu, Chuanxu Feng, Xiaohui Yuan, Lin Zhou, Wenbin Wang, Jie Qin, and Zhongwen Luo. Clip-aware expressive feature learning for video-based facial expression recognition. *Information Sciences*, 598:182–195, 2022. 8

[58] Yuanyuan Liu, Wenbin Wang, Chuanxu Feng, Haoyu Zhang, Zhe Chen, and Yibing Zhan. Expression snippet transformer for robust video-based facial expression recognition. *Pattern Recognition*, 138:109368, 2023. 8

[59] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 8

[60] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5): e0196391, 2018. 2, 3, 4

[61] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3, 4

[62] Fuyan Ma, Bin Sun, and Shutao Li. Spatio-temporal transformer for dynamic facial expression recognition in the wild. *arXiv preprint arXiv:2205.04749*, 2022. 8

[63] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*, 2023. 8

[64] Himangi Mittal, Pedro Morgado, Unnat Jain, and Abhinav Gupta. Learning state-aware visual representations from audible interactions. *Advances in Neural Information Processing Systems*, 35:23765–23779, 2022. 8

[65] Xuan-Bac Nguyen, Chi Nhan Duong, Xin Li, Susan Gauch, Han-Seok Seo, and Khoa Luu. Micron-bert: Bert-based facial micro-expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1482–1492, 2023. 3, 9

[66] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 8

[67] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015. 8

[68] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 8

[69] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 8

[70] Vandana Rajan, Alessio Brutti, and Andrea 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)*, pages 4693–4697. IEEE, 2022. 6, 7

[71] Samir Sadok. *Audiovisual speech representation learning applied to emotion recognition*. PhD thesis, CentraleSupélec, 2024. 6, 9

[72] Pritam Sarkar, Aaron Posen, and Ali Etemad. Avcaffe: a large scale audio-visual dataset of cognitive load and affect for remote work. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 76–85, 2023. 2, 4, 8

[73] Abhronil Sengupta, Yuting Ye, Robert Wang, Chiao Liu, and Kaushik Roy. Going deeper in spiking neural networks: Vgg and residual architectures. *Frontiers in neuroscience*, 13:95, 2019. 9

[74] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations*, 2022. 7, 8

[75] Lang Su, Chuqing Hu, Guofa Li, and Dongpu Cao. Msaf: Multimodal split attention fusion. *arXiv preprint arXiv:2012.07175*, 2020. 9

[76] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Maedfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6110–6121, 2023. 2, 6, 7, 8, 9

[77] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Hicmae: Hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition. *Information Fusion*, 108:102382, 2024. 2, 3, 6, 7, 8, 9

[78] Licai Sun, Zheng Lian, Kexin Wang, Yu He, Mingyu Xu, Haiyang Sun, Bin Liu, and Jianhua Tao. Svfap: Self-supervised video facial affect perceiver. *IEEE Transactions on Affective Computing*, 2024. 6, 7, 8, 9

[79] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 4

[80] Zeng Tao, Yan Wang, Junxiong Lin, Haoran Wang, Xinji Mai, Jiawen Yu, Xuan Tong, Ziheng Zhou, Shaoqi Yan, Qing Zhao, et al. A ˆ{3} lign-dfer: Pioneering comprehensive dynamic affective alignment for dynamic facial expression recognition with clip. *arXiv preprint arXiv:2403.04294*, 2024. 7, 8

[81] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 3, 8

[82] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 6, 7, 8

[83] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 8

[84] Minh Tran and Mohammad Soleymani. A pre-trained audio-visual transformer for emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4698–4702. IEEE, 2022. 6, 7

[85] Minh Tran, Yelin Kim, Che-Chun Su, Cheng-Hao Kuo, and Mohammad Soleymani. Saaml: A framework for semi-supervised affective adaptation via metric learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6004–6015, 2023. 2, 7

[86] 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*, page 6558. NIH Public Access, 2019. 9

[87] Yuan Tseng, Layne Berry, Yi-Ting Chen, I-Hsiang Chiu, Hsuan-Hao Lin, Max Liu, Puyuan Peng, Yi-Jen Shih, Hung-Yu Wang, Haibin Wu, et al. Av-superb: A multi-task evaluation benchmark for audio-visual representation models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6890–6894. IEEE, 2024. 8

[88] Nguyen Van Quang, Jinhee Chun, and Takeshi Tokuyama. Capsulenet for micro-expression recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019. 9

[89] Sergey Verbitskiy, Vladimir Berikov, and Viacheslav Vyshegorodtsev. Eranns: Efficient residual audio neural networks for audio pattern recognition. *Pattern Recognition Letters*, 161:38–44, 2022. 9

[90] Hanyang Wang, Bo Li, Shuang Wu, Siyuan Shen, Feng Liu, Shouhong Ding, and Aimin Zhou. Rethinking the learning paradigm for dynamic facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17958–17968, 2023. 8

[91] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023. 3

[92] Yan Wang, Yixuan Sun, Wei Song, Shuyong Gao, Yiwen Huang, Zhaoyu Chen, Weifeng Ge, and Wenqiang Zhang. Dpcnet: Dual path multi-excitation collaborative network for facial expression representation learning in videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 101–110, 2022. 8

[93] Zhifeng Wang, Kaihao Zhang, Wenhan Luo, and Ramesh Sankaranarayana. Htnet for micro-expression recognition. *Neurocomputing*, 602:128196, 2024. 9

[94] Zhaoqiang Xia, Wei Peng, Huai-Qian Khor, Xiaoyi Feng, and Guoying Zhao. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. *IEEE Transactions on Image Processing*, 29: 8590–8605, 2020. 9

[95] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. Casme ii:

An improved spontaneous micro-expression database and the baseline evaluation. *PloS one*, 9(1):e86041, 2014. 2

[96] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. Multimodal speech emotion recognition using audio and text. In *2018 IEEE spoken language technology workshop (SLT)*, pages 112–118. IEEE, 2018. 6, 7

[97] Seunghyun Yoon, Subhadeep Dey, Hwanhee Lee, and Kyomin Jung. Attentive modality hopping mechanism for speech emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3362–3366. IEEE, 2020. 6, 7, 8

[98] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017. 6

[99] Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6182–6186. IEEE, 2022. 9

[100] Hongyi Zhang. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 4

[101] He Zhang and Hanling Zhang. A review of micro-expression recognition based on deep learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 01–08. IEEE, 2022. 9

[102] Kejun Zhang, Xinda Wu, Xinhang Xie, Xiaoran Zhang, Hui Zhang, Xiaoyu Chen, and Lingyun Sun. Werewolf-xl: A database for identifying spontaneous affect in large competitive group interactions. *IEEE Transactions on Affective Computing*, 14(2):1201–1214, 2021. 2, 4, 8

[103] Liangfei Zhang, Xiaopeng Hong, Ognjen Arandjelović, and Guoying Zhao. Short and long range relation based spatio-temporal transformer for micro-expression recognition. *IEEE Transactions on Affective Computing*, 13(4): 1973–1985, 2022. 9

[104] Xiaoqin Zhang, Min Li, Sheng Lin, Hang Xu, and Guobao Xiao. Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 6, 7, 8

[105] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):915–928, 2007. 9

[106] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1553–1561, 2021. 6, 7, 8

[107] Zengqun Zhao and Ioannis Patras. Prompting visual-language models for dynamic facial expression recognition. *arXiv preprint arXiv:2308.13382*, 2023. 7, 8

[108] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30:6544–6556, 2021. 8, 9

[109] Ling Zhou, Qirong Mao, and Luoyang Xue. Dual-inception network for cross-database micro-expression recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–5. IEEE, 2019. 9

[110] Ling Zhou, Qirong Mao, Xiaohua Huang, Feifei Zhang, and Zhihong Zhang. Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition. *Pattern Recognition*, 122:108275, 2022. 3, 9

[111] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pages 650–667. Springer, 2022. 1, 2