2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Social-MAE: A Transformer-Based Multimodal Autoencoder for Face and Voice

Hugo Bohy[1], Minh Tran[2], Kevin El Haddad[1], Thierry Dutoit[1] and Mohammad Soleymani[2]
[1] Numediart Institute, ISIA Lab, University of Mons, Mons, Belgium
[2] Institute for Creative Technologies, University of Southern California, Los Angeles, CA, USA

Abstract— **Human social behaviors are inherently multimodal necessitating the development of powerful audiovisual models for their perception. In this paper, we present Social-MAE, our pre-trained audiovisual Masked Autoencoder based on an extended version of Contrastive Audio-Visual Masked Auto-Encoder (CAV-MAE), which is pre-trained on audiovisual social data. Specifically, we modify CAV-MAE to receive a larger number of frames as input and pre-train it on a large dataset of human social interaction (VoxCeleb2) in a self-supervised manner. We demonstrate the effectiveness of this model by fine-tuning and evaluating the model on different social and affective downstream tasks, namely, emotion recognition, laughter detection and apparent personality estimation. The model achieves state-of-the-art results on multimodal emotion recognition and laughter recognition and competitive results for apparent personality estimation, demonstrating the effectiveness of in-domain self-supervised pre-training. Code and model weight are available here https://github.com/HuBohy/SocialMAE.**

I. INTRODUCTION

Human emotions and social behaviors are expressed and perceived through multiple modalities. While verbal communication can provide information on a person's communicative intent and emotions, non-verbal communication has shown to be equally or even more important [26]. Socially intelligent systems require multimodal methods allowing them to perceive human social and expressive behaviors. Understanding expressions and social behaviors can be achieved by analyzing audiovisual modalities, i.e., face, body and voice. Although unimodal approaches, e.g., vision from facial expression or audio for tracking arousal, can reach a high performance [5], [18], fusing two modalities increases the efficiency and robustness of multimodal systems [29], [10]. Information from different modalities can be congruent and reinforce each other for more effective communication. For instance, a smiling face aligned with a cheerful tone reinforce the expression of happiness. Information from different modalities can also be complementary, improving clarity and reducing uncertainty, such as a confident tone coupled with a smiling face. There is also a possibility of interaction between modalities generating new meanings from contradictory information from different modalities, e.g., expression of irony with conflicting face and voice behaviors.

In supervised learning, the availability of labeled data is often limited by laborious annotation. A common solution is to fine-tune a pre-trained model, i.e., to use models that have previously been trained on a larger dataset of similar nature [8], [35], [22]. Self-supervised learning has been proposed to leverage large-scale unlabeled datasets to pre-train a model by pre-training models using a pretext task. One such pretext task is autoencoding, i.e., encoding input information into an often more compact latent representation and decoding it back to the original space. The encoder module of an autoencoder can be re-used as a pre-trained model with powerful and discriminative representations to be fine-tuned or re-used for downstream tasks.

Past work extensively explored the natural interactions between audio and visual signals for representation learning [32], [33], [31], [30], [1], [11], [37] through self-supervision with a variety of pretext tasks. Synthesis-based strategies [37], [32], [33] have been proposed, where audio and visual signals are artificially combined to facilitate learning cross-modal associations. Alignment-based methods [27], [31], [2], [12], on the other hand, focus on aligning signals from both modalities in time or space, aiming to extract meaningful correlations between them. Another line of research involves the application of masked autoencoding (MAE) [14], where the model learns to reconstruct the missing portions of either the audio or visual input, fostering representation learning through learning the structure of the data. Recently, two models, namely MAViL [25] and CAV-MAE [20], have explored the combination of MAE with contrastive learning and demonstrated state-of-the-art (SOTA) performance on audio-visual classification. Adding contrastive learning allows the models to learn inter-modality representations.

Despite the popularity of emotion and social behavior perception, datasets for such tasks are often limited in size due to the high cost of labeling. Most existing audiovisual methods are based either on transfer learning with models trained on out-of-domain data, e.g., AudioSet [13], or trained from scratch. However, the desired input data should contain human faces and voices. Existing audiovisual encoders, e.g., [19], also lack the temporal fidelity in the visual domain. In contrast, expressive behaviors in the human face are rather dynamic and fast-moving. There is limited work, e.g., [38], on audiovisual encoders suitable for the automatic perception of human emotional and communicative behaviors.

The work was partially sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-20-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

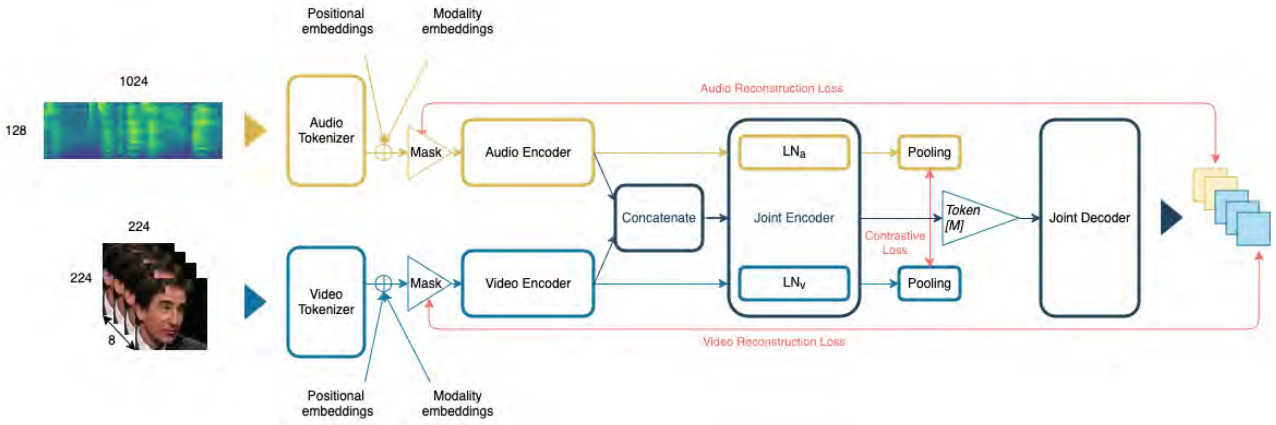


Fig. 1: Social-MAE model for voice and face analysis in videos. The model is pre-trained to reconstruct audio and visual modalities from masked portions of their corresponding input, narrowing the difference between each modality representation.

In this paper, we present Social-MAE, a pre-trained audiovisual model based on Masked Autoencoder. We aim to adapt a self-supervised method with superior results on audio event recognition for the audiovisual understanding of human social behaviors. We evaluate our model against several baselines on three different social and affective tasks: emotion recognition, laughter detection and apparent personality estimation. The main contributions of this work are as follows.

- We present Social-MAE, a model based on CAV-MAE architecture adapted to social context by pre-training on a large-scale social dataset;
- To develop Social-MAE, we modify CAV-MAE to accept multiple frames providing higher temporal fidelity at visual input;
- Our experiments demonstrate the importance of in-domain pre-training for affective and social tasks. Our model reaches or outperforms SOTA models on relevant tasks.

II. METHOD

In this section, we present Social-MAE (Fig. 1), an adapted version of CAV-MAE that focuses on voice and face. The model is composed of two modality-specific encoders followed by a joint encoder module and a joint decoder module. Each module relies on a set of Transformer layers [39] made of an attention block, a feed-forward network, residual connections and layer normalization [3]. We describe the pre-processing pipeline in Sec. II-A, the model overview in Sec. II-B and the self-supervised training in Sec. II-C.

A. Audiovisual Tokenization

The architecture follows a mid-fusion scheme: both audio and video are first encoded in two separate branches for several encoder layers before merging into a joint encoder. Audio data are pre-processed as in CAV-MAE: we convert the input audio waveform into a sequence of 128-dimensional log Mel filterbank features computed with a 25 ms Hamming window and an overlap of 10 ms. We pad or crop the length of the input to keep 1024 audio frames, resulting in a 128×1024 spectrogram. The spectrogram is processed as an image that we split in N 16×16 non-overlapping patches. Each patch is projected with a linear layer to a 1-dimensional embedding of size 768, referred to as a *token*. We add a trainable positional embedding to each token to provide information about the token order.

Visual inputs differ from CAV-MAE as they consist of eight randomly selected frames as proposed in [25] rather than single frame. Each frame is an RGB image of the face bounding box scaled to 224×224 pixels, resulting in a $8 \times 224 \times 224 \times 3$ video input. We split the video into N $2 \times 16 \times 16$ patches with no overlap, flatten and project with a linear layer into tokens of size 768. A trainable positional embedding is added to each token as well. Another trainable parameter provides information about each token's modality and weights the modality's importance. After adding positional and modality embeddings, a random mask with a rate of p% is applied to the input tokens, providing the model only with (1-p)% of the original audio and/or video sequence.

B. Model Description

This section presents an overview of the autoencoder architecture as described in [20]. The model first processes an input sequence in separate encoders, each leveraging unimodal information. The modality encoders are stacks of 11 Transformer layers that aim to encode internal patterns in the input sequence. The joint encoder comprises a single Transformer layer on top of the modality encoders. Each modality is processed by the respective encoder followed by the joint encoder either individually or concatenated with the second modality depending on the targeted loss. The layer normalization on top of the joint encoder differs for audio, video and multimodal processing. The weights of the joint encoder are shared regardless of its input modality, as it was shown that weight sharing lightens the model without degrading performance [28]. The unimodal tokens are averaged following the average pooling method, while the multimodal tokens are fed to the joint decoder, which is a stack of 8 Transformer layers. It aims to retrieve the original video and audio from an input sequence made of the encoded tokens and a learnable token **M** repeated at masked positions. The reconstruction loss is the distance between tokens **M** at the output of the decoder and their corresponding original tokens.

C. Self-Supervised Pre-training

We adapted the pre-trained CAV-MAE model by training with self-supervision on the VoxCeleb2 dataset [7]. VoxCeleb2 is an audiovisual dataset that contains over a million utterances from more than 6,000 speakers of 145 different nationalities. It provides a wide range of languages, accents, ethnicities and ages from real-world recordings. As self-supervised pre-training often requires vast amounts of data, we chose VoxCeleb2, as a suitable large and diverse audiovisual dataset with social content.

The learning phase relies on the weighted combination of contrastive and reconstruction loss that provides complementary information. For an input sequence of N pairs of audio and video tokens a_i, v_i, the contrastive loss $\mathcal{L}_c$ is computed on modality averaged tokens c_i^a, c_i^v and aims to leverage relevant inter-modal information by following a LogSoftmax loss. The reconstruction loss $\mathcal{L}_r$ evaluates the model's ability to reconstruct masked tokens x_i^{mask} from the tokens at the output of the decoder $\widehat{M_i}$ with an MSE loss. The final loss is the weighted sum of the contrastive and the reconstruction losses: $\mathcal{L} = \mathcal{L}_c \cdot \lambda_c + \mathcal{L}_r$.

III. EXPERIMENTS AND RESULTS

We pre-trained our Social-MAE during 25 epochs with a learning rate starting at 10^{-4} and decreasing at a decay rate of 0.5 every 5 epochs with a masking ratio p=75%. For comparison, we also pre-trained CAV-MAE (as it uses 1 frame instead of 8) following the same settings. Both models were initialized on CAV-MAE^{scale+} weights pre-trained on AudioSet-2M with self-supervision. We report visual zero-shot reconstruction in Fig. 2 using pre-trained Social-MAE on two downstream task datasets: CREMA-D [6] and ChaLearn First Impressions (FI) [36]. The model is able to provide a convincing output on previously unseen data. Most reconstruction errors, although not obvious at first sight, come from the most dynamic areas of the face, such as the eyes or lips.

For downstream tasks, we remove the decoder from the architecture and replace it with a randomly initialized linear layer. We evaluate our pre-trained model by fine-tuning it on three different social and affective tasks: emotions recognition on CREMA-D [6], personality traits regression on ChaLearn FI [36] and smiles and laughter detection on NDC-ME [23]. For each task, we describe the dataset, the fine-tuning pipeline and the evaluation metrics to compare CAV-MAE and Social-MAE models against published baselines, following their experimental settings for consistency.

A. Emotion Recognition

1) Experimental setup: This task is evaluated on the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D), containing 7,442 clips from 48 male and 43 female actors (20-74 years old). Each actor spoke 12 sentences using one of six emotions (anger, disgust, fear, happiness, sadness and neutral) ranging from 763 to 2204 utterances per emotion. Fine-tuning requires no masking on audio and visual tokens. We fine-tune pre-trained Social-MAE as well as our pre-trained version of CAV-MAE for 20 epochs using a mini-batch size of 8, learning rates at 10^{-4} and 10^{-5} for the encoders and the head respectively and we use the Cross-Entropy Loss.

2) Baselines:

a) UAVM: [19] presented UAVM, a unified audiovisual framework for classification. The model uses pre-trained CNN-based feature extractors on log Mel filterbanks and multi-frame visual inputs that are fed to Transformer layers.

b) AuxFormer: [16] proposed AuxFormer, a multimodal model that fuses audio and visual tokens through Transformer inputs. The model also processes separate modalities through auxiliary networks. The model loss is a weighted combination of the network losses. Audio inputs are low-level descriptors from OpenSmile [9] toolkit, and visual inputs are face clips processed by pre-trained VGG-face architecture [34].

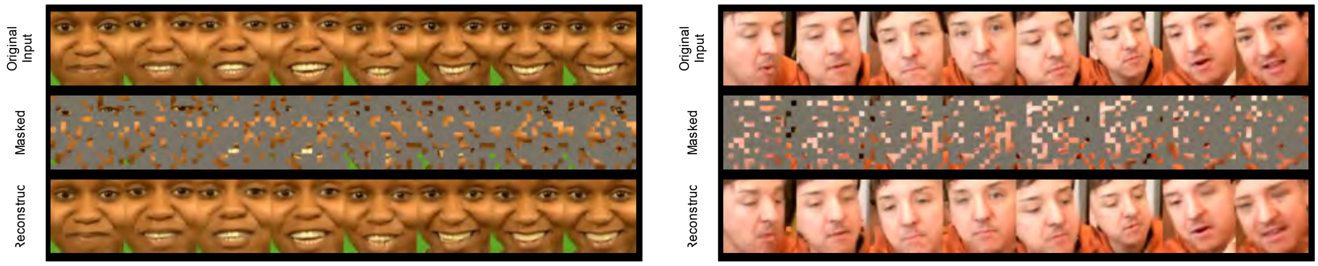
c) VAVL: [17] proposed an audiovisual model, named Versatile AudioVisual Learning (VAVL), which relies on the Conformer architecture [21]. Each modality input flows through a separate encoder followed by a shared-weight conformer. Audio inputs are high-dimensional features from Wav2vec2.0 [24] and visual inputs are face clips processed into emotional feature representations.

TABLE I: F1-score performance Comparison on CREMA-D. Mi and Ma refer to F1-score Micro and Macro. The best results are in **bold** face font. * p-value < 1E-5

	Audio		**Visual**		**AV**	
	Mi	**Ma**	**Mi**	**Ma**	**Mi**	**Ma**
AuxFormer [16]	0.648	0.593	0.626	0.560	0.763	0.698
UAVM [19]	0.554	0.614	0.672	0.617	0.769	0.749
VAVL [17]	**0.701**	0.628	**0.787**	0.738	0.826	0.779
CAV-MAE	0.694	**0.694**	0.630	0.635	0.766	0.759
Social-MAE	0.601*	0.607*	0.749*	**0.755***	**0.837***	**0.842***

3) Results and discussion: Table I reports the F1-score with micro and macro averaging techniques. Social-MAE outperforms previously published methods for audiovisual classification. The micro F1 score shows the global accuracy, and the macro F1 score shows the unweighted average accuracy across each class, so the macro F1 score can be influenced by class imbalance. Since classes in CREMA-D range from 763 utterances (Sadness) to 2204 utterances (Neutral), we interpret the similarities between the macro and micro F1-scores reached by our pre-trained models as their ability to recognize emotions regardless of their prevalence.

Adapted CAV-MAE competes for best audio-only classification against VAVL model. Social-MAE rivals the best baseline for visual classification. The performance is impressive when you consider that the former processes 8 frames and the latter processes high-level features from all input frames. We also find it interesting that adapted CAV-MAE is able to outperform multi-frame baselines AuxFormer and UAVM on



(a) Reconstruction on CREMA-D.

(b) Reconstruction on ChaLearn First Impressions.

Fig. 2: Social-MAE visual zero-shot reconstruction on CREMA-D and ChaLearn First Impressions datasets. The first row shows the original input, the second row the visual equivalent to masked tokens, and the last row the reconstructed frames.

both unimodal and multimodal classification tasks, highlighting the efficiency of in-domain self-supervised pre-training.

B. *Personality Trait Prediction*

1) Experimental setup: We evaluate personliaty prediction with the First Impressions (FI) dataset, a collection of 10,000 *in-the-wild* videos, in average 15s long. Videos are annotated with apparent personality traits known as *big-5* [15]: Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism. Fine-tuning requires no masking on audio and visual tokens. We fine-tuned both models presented in Sec. II-C for 10 epochs using a mini-batch size of 8, an encoder learning rate of 1e-4 and the classification head learning rate of 1e-5. We use a Mean Absolute Error loss and our accuracy metric is 1 − Mean Absolute Error.

We compare our fine-tuned CAV-MAE and Social-MAE to the best team of the challenge associated to the dataset: NJU-LAMDA [40], a model pre-trained on VGG-face. The audio input is log Mel filterbank and the visual input is the deep features from 100 frames. They train their model in 100 epochs for the audio stream and 3 epochs for the pre-trained visual stream, with a mini-batch of 128.

TABLE II: Model Accuracy on First Impressions Dataset. Best results are in **bold** face font. * p-value $<$ 1E-5.

	Ope.	Con.	Ext.	Agr.	Neu.	Avg.
NJU-LAMDA [40]	**0.912**	**0.916**	**0.913**	**0.913**	**0.910**	**0.913**
CAV-MAE	0.899	0.899	0.899	0.902	0.896	0.899
Social-MAE	0.908*	0.902*	0.895*	0.907*	0.905*	0.903*

2) Results and discussion: Table II shows the accuracy of each personality trait on ChaLearn First Impressions dataset as well as the mean accuracy. Social-MAE shows a performance of 90.32% on average. While the accuracy is lower than the baseline, it remains impressive considering it was trained for only 10 epochs and with a smaller mini-batch size. We can also observe that processing multiple frames simultaneously (Social-MAE) demonstrates better regressions on four out of five traits compared to the single frame method (CAV-MAE).

C. *Smiles and Laughter Detection*

1) Experimental setup: The Naturalistic Dyadic Conversation on Moral Emotions (NDC-ME) dataset contains 8,352 clips of interactions in English of participants from different backgrounds. Each clip lasts 1.22 seconds, is cropped around the face, and is annotated with non-verbal expressions of smile, laughter, and neutral. We fine-tuned for 10 epochs, with no masking strategy, a mini-batch of 8, and learning rates of 1e-5 and 1e-4 for the backbone and classification head, respectively. Our training objective is the Cross-Entropy Loss. The baseline for smile and laughter detection is LSN-TCN [4], a CNN-based architecture that processes embedded representations of audio and video input separately and feeds them to two fully-connected joint layers.

TABLE III: F1-score on NDC-ME. Best results are in **bold** face font. * p-value $<$ 1E-5.

	Pre-training	Audio	Visual	Audiovisual
LSN-TCN [4]	Supervised	0.438	0.608	0.590
CAV-MAE	Self-Supervised	0.471	0.629	0.766
Social-MAE	Self-Supervised	**0.546***	**0.728***	**0.776***

2) Results and discussion: Table III shows that both self-supervised methods reach higher F1-scores than the supervised baseline. Using multiple frames instead of one significantly improves the performance of the visual modality while slightly improving that of the multimodal classification.

IV. CONCLUSIONS

In this paper, we presented Social-MAE, our pre-trained audiovisual Masked AutoEncoder on audiovisual social data. We modified existing CAV-MAE to accept multiple frames on a large human social behavior dataset. We evaluated our model on three relevant downstream tasks, demonstrating its effectiveness in achieving state-of-the-art results in audiovisual emotion recognition with a 0.837 F1 score and laughter detection with a 0.776 F1 score. With this work, we demonstrated the significance of in-domain adaptation of a large multimodal model trained through self-supervised pre-training. The proposed pre-trained encoder can be easily fine-tuned for other audiovisual social behavior understanding tasks, enabling more robust and performant models for perceiving human behaviors.

References

[1] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman. Self-supervised learning of audio-visual objects from video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 208–224. Springer, 2020.

[2] R. Arandjelovic and A. Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017.

[3] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[4] H. Bohy, K. El Haddad, and T. Dutoit. A New Perspective on Smiling and Laughter Detection: Intensity Levels Matter. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8.

[5] Z. Cai, S. Ghosh, K. Stefanov, A. Dhall, J. Cai, H. Rezatofighi, R. Haffari, and M. Hayat. Marlin: Masked autoencoder for facial video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1493–1504, 2023.

[6] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

[7] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 1459–1462. Association for Computing Machinery.

[10] H. M. Fayek and A. Kumar. Large scale audiovisual learning of sounds with weakly labeled data. *arXiv preprint arXiv:2006.01595*, 2020.

[11] A. Furnari and G. M. Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4021–4036, 2020.

[12] R. Gao, C. Chen, Z. Al-Halah, C. Schissler, and K. Grauman. Visualechoes: Spatial image representation learning through echolocation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 658–676. Springer, 2020.

[13] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE.

[14] M.-I. Georgescu, E. Fonseca, R. T. Ionescu, M. Lucic, C. Schmid, and A. Arnab. Audiovisual masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16144–16154, 2023.

[15] L. R. Goldberg. An Alternative "Description of Personality": The Big-Five Factor Structure. In *Personality and Personality Disorders*. Routledge.

[16] L. Goncalves and C. 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.

[17] L. Goncalves, S.-G. Leem, W.-C. Lin, B. Sisman, and C. Busso. Versatile audio-visual learning for handling single and multi modalities in emotion regression and classification tasks. *arXiv preprint arXiv:2305.07216*, 2023.

[18] Y. Gong, Y.-A. Chung, and J. Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

[19] Y. Gong, A. H. Liu, A. Rouditchenko, and J. Glass. Uavm: Towards unifying audio and visual models. *IEEE Signal Processing Letters*, 29:2437–2441, 2022.

[20] Y. Gong, A. Rouditchenko, A. H. Liu, D. Harwath, L. Karlinsky, H. Kuehne, and J. Glass. Contrastive audio-visual masked autoencoder. *arXiv preprint arXiv:2210.07839*, 2022.

[21] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[22] A. Hassani, S. Walton, J. Li, S. Li, and H. Shi. Neighborhood Attention Transformer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6185–6194. IEEE.

[23] L. Heron, J. Kim, M. Lee, K. El Haddad, S. Dupont, T. Dutoit, and K. Truong. A Dyadic Conversation Dataset on Moral Emotions. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 687–691.

[24] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, et al. Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. *arXiv preprint arXiv:2104.01027*, 2021.

[25] P.-Y. Huang, V. Sharma, H. Xu, C. Ryali, Y. Li, S.-W. Li, G. Ghosh, J. Malik, C. Feichtenhofer, et al. Mavil: Masked audio-video learners. *Advances in Neural Information Processing Systems*, 36, 2024.

[26] M. L. Knapp, J. A. Hall, and T. G. Horgan. *Nonverbal communication in human interaction*, volume 1. Holt, Rinehart and Winston New York, 1978.

[27] B. Korbar, D. Tran, and L. Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems*, 31, 2018.

[28] S. Lee, Y. Yu, G. Kim, T. Breuel, J. Kautz, and Y. Song. Parameter efficient multimodal transformers for video representation learning. *arXiv preprint arXiv:2012.04124*, 2020.

[29] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34:14200–14213, 2021.

[30] E. Ng, D. Xiang, H. Joo, and K. Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9890–9900, 2020.

[31] A. Owens and A. A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 631–648, 2018.

[32] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2405–2413, 2016.

[33] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Ambient sound provides supervision for visual learning. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 801–816. Springer, 2016.

[34] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep Face Recognition. In *Procedings of the British Machine Vision Conference 2015*, pages 41.1–41.12. British Machine Vision Association.

[35] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[36] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera. ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results. In G. Hua and H. Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, volume 9915, pages 400–418. Springer International Publishing.

[37] B. Shi, A. Mohamed, and W.-N. Hsu. Learning lip-based audio-visual speaker embeddings with av-hubert. *arXiv preprint arXiv:2205.07180*, 2022.

[38] M. Tran, Y. Kim, C.-C. Su, C.-H. Kuo, and M. Soleymani. Saaml: A framework for semi-supervised affective adaptation via metric learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 6004–6015, New York, NY, USA, 2023. Association for Computing Machinery.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[40] C.-L. Zhang, H. Zhang, X.-S. Wei, and J. Wu. Deep Bimodal Regression for Apparent Personality Analysis. In G. Hua and H. Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, Lecture Notes in Computer Science, pages 311–324. Springer International Publishing.