Tackling Missing Modalities in Audio-Visual Representation Learning Using Masked Autoencoders

Georgios Chochlakis[1,*], *Chandrashekhar Lavania*[2], *Prashant Mathur*[2], *Kyu J. Han*[2]

[1]University of Southern California, USA
[2]AWS AI Labs, USA

chochlak@usc.edu, clavania@amazon.com, pramathu@amazon.com, kyujhan@amazon.com

Abstract

Audio-visual representations leverage information from both modalities to produce joint representations. Such representations have demonstrated their usefulness in a variety of tasks. However, both modalities incorporated in the learned model might not necessarily be present all the time during inference. In this work, we study whether and how we can make existing models, trained under pristine conditions, robust to *partial* modality loss without retraining them. We propose to use a curriculum trained Masked AutoEncoder, to impute features of missing input segments. We show that fine-tuning of classification heads with the imputed features makes the base models robust on multiple downstream tasks like emotion recognition and Lombard speech recognition. Among the 12 cases evaluated, our method outperforms strong baselines in 10 instances.
Index Terms: video, speech, masked autoencoder, missing modality

1. Introduction

Recent works on multimodal audio-visual representation learning have been successful on a variety of tasks like action recognition [1] and audio-visual speech recognition [2]. A typical multimodal model learns representations of speech and video separately and projects them in a common space [1, 3]; this allows the model to leverage potentially supplementary unimodal information. However, there are scenarios where segments from either or both modalities are missing or dropped out, because of low bandwidth, corrupted signals, etc. in communication systems in practice. These multimodal models, if not robust to such dropouts, can suffer from a catastrophic regression in performance with a partial or complete loss of one modality [4, 2]. Moreover, if the performance of a multimodal model degrades below that of unimodal ones, there is no reason to use the multimodal model in the first place. This enforces a strong need to build a multimodal model that is robust to this dropout scenario.

In some recent works, modality loss has been dealt via unimodal back-offs [2, 5], which can take the form of a cascade of models or a mixture of experts (MoEs). Other approaches have focused on training with a simulation of the test scenario [2]. In this paper, we propose *feature imputation* to improve audio-visual multimodal models' robustness to partial loss of both speech and video input sequences without retraining the models from scratch, which would save training time as well as infrastructure cost to support the model training as multimodal models employing neural network architectures get larger. The proposed method makes existing multimodal models robust, without retraining them, by interjecting an imputation module based on a Masked AutoEncoder (MAE) that estimates feature values for dropped out clips instead of zero-ing them [6]. MAEs have been introduced in various settings, like vision [7], natural language [8], and audio-visual retrieval or event classification tasks [9]. They are trained to reconstruct parts of inputs that have been artificially masked for learning robust representations of the input in the process. Thereafter, they are typically used as feature extractors in downstream tasks [8, 7, 10]. However, their training regime of masked (or missing) inputs is appropriate for our inference setting, and indeed we show that MAEs can also be robustly used to impute missing features for modality dropout even with frozen backbone feature extractors.

We break our inputs into small, continuous clips (e.g, 500 msec duration), and consider modality dropouts at that level instead of individual frames. This is a more realistic setting than individual missing frames. For instance, a network packet may contain several audio/visual frames, thus its loss will result in losing a clip instead of a frame. We focus on cases where either or both modalities can be dropped out. As the inference dropout rate is not expected to be fixed, we propose to train the MAE-based feature imputation module with a curriculum of increasing dropout rates. We perform experiments on sequence classification tasks, like CREMA-D [11], RAVDESS [12], and Lombard speech recognition [13]. Our contributions include

1. An audio-visual MAE to impute missing input segments during inference for an existing classifier. To the best of our knowledge, this is the first work to use imputation with MAEs to improve robustness to modality dropouts.
2. A plug-and-play approach that can be integrated with any classifier and is 14x faster than (re-)training of major parts.
3. Comprehensive experiments on three different data sets covering a wide range of dropout levels for both modalities and studying how modality drop affects downstream tasks.

2. Related Work

2.1. Robustness to modality dropout

As multimodal training is actively researched in the present [14, 15], studying the impact of modality dropout also receives needed attention. Many approaches to modality dropout in audio-visual settings consider the case where one of the modalities is entirely missing [16, 17]. Others have developed strategies for robustness to missing input segments, such as training a multimodal model with randomly omitted audio/visual frames as well as dropping an entire modality [18, 19, 20, 2]. In this case, the classifier is trained to handle missing segments by itself, making the distribution shift during testing less severe. Other techniques opt for unimodal backoffs, such as mixture of experts (MoEs) and model cascades [2, 5], wherein the components of a model that will process the input are selected

[*]Work done as intern at Amazon; not a corresponding author

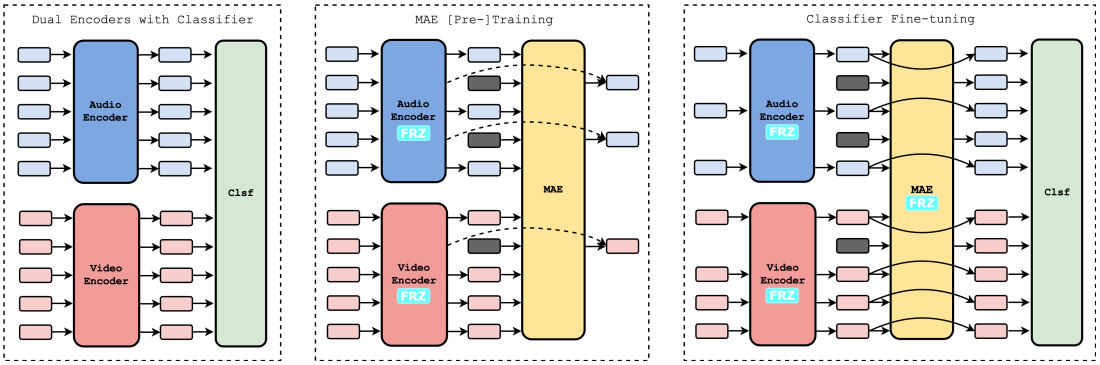


Figure 1: *Different training stages in our proposed approach: 1) standard encoder training with a classifier head, 2) fine-tuning of MAE with the frozen encoders to reconstruct missing features with a curriculum of varying masking dropout rates and 3) fine-tuning of the classifier head with the output of MAE.*

based on which modality is available at inference time. These methods either require retraining of multimodal models from scratch, come with carrying out multiple models (in the form of additional unimodal models), or avoid explicitly modeling the missing modality setting, but rather work around it.

2.2. Imputation

Our focus in this paper is time-series data that contains missing segments (i.e., a collection of contiguous frames), so we limit the discussion in this section to this use case. Autoencoders have been used successfully for imputation or data cleaning, often in the form of Denoising or Variational Autoencoders (VAEs). For example, [21] propose a partial VAE that utilizes the conditional independence between observed and missing data given the latent variables. [22] introduce the use of an input dropout distribution during the training of a VAE. We, on the other hand, re-purpose MAE to perform the imputation. MAEs have been used in natural language processing [8] and vision [7] as well as for audio-visual representations learning [9, 23]. CAV-MAE [9] uses a multimodal encoder-decoder architecture, where the masked tokens are introduced in the decoder. The output embeddings produced by the encoder are also guided by a cross-modal contrastive loss. Note that the masking in CAV-MAE is performed on image and spectrogram *patches*, whereas we mask *sequences* of audio and visual frames in the paper. While being used in MAEs to learn good representations during training, the masked regime in our case is for inference to impute missing feature sequences. The proposed MAE-based approach for feature imputation can be considered as a more natural alternative to generative models such as GANs [24], which can be used to impute missing values, both in the form of the raw signal itself and of speech or visual features [25].

3. Methodology

3.1. Approach

We propose to use MAEs during inference to perform imputation of missing segments at the feature level in an audio-visual setup (Figure 1). We reconstruct *only* missing inputs, and the final encoding with imputed features is fed to the classification head. We fine-tune the latter to alleviate the shift between real and imputed features. Since we handle modality inputs chuck-wise, we leverage intermediate representations for those chunks from the encoders per modality. We assume that our visual chunk unit is clips rather than individual frames. This allows us to impute features of the entire clip. This enables us to leverage visual encoders that integrate temporal information, like AVID-CMA [1] using a spatiotemporal convolution block called R(2+1)D [26], which prevents the missing inputs from "contaminating" neighboring frames [27, 28]. For audio modality, we use a similar chunking strategy, as dropping out individual frames is not a major concern due to the high sampling frequency of the domain and audio encoders usually rely on temporal information beyond very small windows [29, 30].

3.2. Backbone Encoders

AVID-CMA [1]: AVID-CMA relies on an R(2+1)D network that provides clip-level features for visual modality. For audio modality, a CNN operates on the log spectrogram of audio. We modify the existing implementation and leverage the pre-trained models provided by the authors.[1] We switch to multimodal evaluation ([1] had unimodal) via feature concatenation.
CrissCross [3]: CrissCross utilizes similar backbones. Apart from the alternative proxy task used for self-supervision, the inference setting is virtually identical to AVID-CMA. We also leverage their existing implementation and pre-trained models.[2]

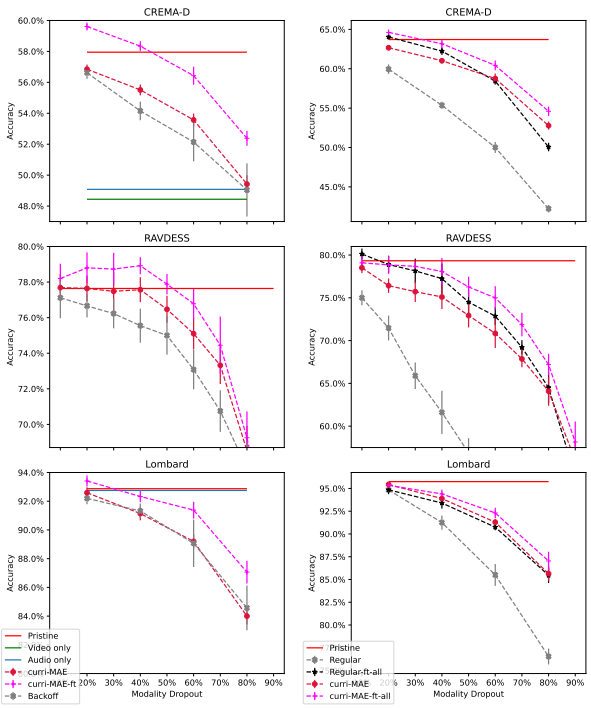
3.3. Curriculum Masked Autoencoder

We find that initializing our MAE from the weights of CAV-MAE [9] results in better downstream performance, even though it uses image and spectrogram patches as opposed to clip-level features. We, therefore, follow their architecture. Because of this misalignment between the domains, we further pre-train the MAE with a fixed dropout rate on AudioSet-20K [31]. Then, we fine-tune the pre-trained MAE on features of the downstream benchmarks with a *curriculum* of increasing masking rates throughout training, from one masked chunk to all but one. In this manner, the MAE-based imputation module is trained in a principled way with variable masking rates, exposing it to the entire gamut of possible inference dropout rates.

3.4. Classifiers

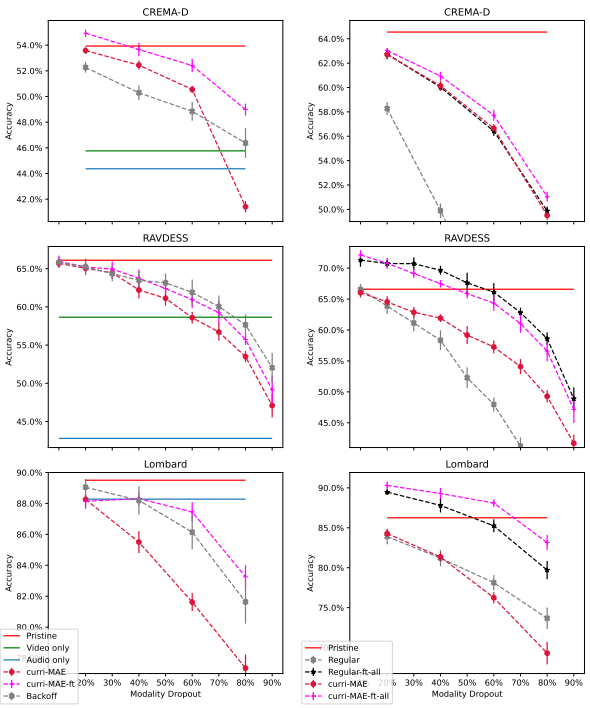
We experiment with pairing the following classifiers with the MAE in our pipeline (shown in Figure 1), and in Section 3.5 we present the corresponding baselines when the MAE is not used.

[1] https://github.com/facebookresearch/AVID-CMA
[2] https://github.com/pritamqu/CrissCross



(a) *MAE+linear and its baseline* (b) *MAE+GRU and its baseline*

Figure 2: *Accuracy of AVID-CMA features at corresponding levels of dropout in both modalities.*



(a) *MAE+linear and its baseline* (b) *MAE+GRU and its baseline*

Figure 3: *Accuracy of CrissCross features at corresponding levels of dropout in both modalities.*

Linear: Following the standard evaluation process of AVID-CMA and CrissCross, we experiment with a linear classifier, independently predicting for individual clips and aggregating the predictions to arrive at a video-level prediction.

GRU [32]: To introduce temporal dependencies, we experiment with treating the output features of the backbone encoders for each clip as a sequence and using a GRU to classify the sequence by using its final hidden representation.

3.5. Baselines

When the MAE is not used, we use the following baselines, each corresponding to the classifiers in Section 3.4:

Unimodal Backoff for Linear Classifier: We train both a multimodal classifier and unimodal classifiers, which we backoff to if one of the modalities is missing. Clips with no modality present are discarded. This backoff approach can be likened to an MoE, and is virtually identical to dropout training with frozen backbones.

Constant Imputation for GRU: Due to the temporal dependency between different clips, there is no obvious backoff to different models depending on the present modalities of clips, nor is it obvious that clips can be skipped when both modalities are absent. Therefore, we replace missing features with a fixed feature vector, and then fine-tune the GRU classifier by randomly dropping segments.

4. Experiments

4.1. Datasets

CREMA-D [11]: 7,442 clips of actors speaking 12 different sentences in one of the following acted emotions: happiness, sadness, anger, fear, disgust, and neutral. We use 70-15-15 speaker-independent random splits.

RAVDESS [12]: 4,230 clips of 24 actors speaking or singing 2 lexically-neutral sentences in one of the following emotions: calm, happy, sad, angry, fearful, surprised and disgusted. We use speaker-independent random splits, with 4 test speakers.

Lombard speech recognition [13]: Lombard speech recognition dataset with labels indicating whether the speech is Lombard [33] or plain. There are 54 talkers with 100 evenly divided utterances each. We refer to this dataset as **Lombard**. We use speaker-independent random splits, with 8 test speakers.

4.2. Implementation Details

We use up to 4 NVIDIA Tesla T4 GPUs for model training. We apply dropout randomly and independently in the two modalities, but keep the dropout ratios *equal* to constrain the possible space of dropouts. We freeze the pre-trained backbone encoders instead of fine-tuning them jointly with the final classifiers. Note that the number of parameters in our MAE (from CAV-MAE) is larger than the number of parameters in the backbone encoders (180M vs. 25M). However, handling video clips makes the throughput of training the backbone encoders considerably worse than the proposed MAE, specifically, 0.7 GPU hrs and 0.05 GPU hrs per epoch on CREMA-D respectively — a 14$\times$ reduction — justifying the choice to keep the backbones frozen. At the same time, initializing weights from CAV-MAE also makes the proposed approach more data-efficient. Our clips are 400ms for CREMA-D and RAVDESS, and 500ms for Lombard and AudioSet, selected heuristically based on video duration statistics. The number of clips we use per video is 10 for Audioset and RAVDESS, and 5 for CREMA-D and Lom-

bard. We found that abandoning the contrastive loss in MAE's training from [9] is preferable as the models converge to a better reconstruction loss. We pre-train the MAE for 50 epochs on AudioSet-20K with a 1e-4 learning rate, and we fix the dropout rate to the closest possible rate to 75% (used in CAV-MAE pre-training). We then fine-tune it on each dataset for 50 epochs, with a 1e-4 learning rate for CREMA-D and Lombard, and 5e-4 for RAVDESS. We increase the dropout rate every 5 epochs in datasets with 10 clips, or every 10 epochs otherwise. We use a 1e-3 learning rate to train the classification heads.

For the constant imputation for the GPU classifier, we found that an all-zeros vector performs favorably in the absence of fine-tuning, and equivalently otherwise to an average feature vector for each modality from AudioSet-20K, hence we present the former. To fine-tune this baseline GRU, for each sample, we randomly sample the level of dropout (which is equal for both modalities) from *all* possible dropout rates, and then sample a dropout mask. We do the same when fine-tuning the GRU with MAE features, whereas we select the largest dropout rate (the last rate in the curriculum) for the fine-tuning of the linear classifier with MAE features.

4.3. Feature Imputation with MAE against Baselines

We present the performance of the imputed features in all datasets. Each data point has equal amounts of dropout in both modalities. We focus on the influence of different levels of dropout and the utilization of different classifiers. Our results are presented in Figure 2 using AVID-CMA, and Figure 3 using CrissCross. We plot the mean and standard dev based on 10 random masks applied to each sample, and we consider the performance of two models equivalent when ranges overlap.

We present *Pristine*, *Video only*, *Audio Only*, *Backoff*, *Regular*, *Regular-ft-all*, *curri-MAE*, *curri-MAE-ft* and *curri-MAE-ft-all*. *Pristine*, *Video Only* and *Audio Only* are the performance of the classifier and unimodal baselines in perfect conditions. *Backoff* is the linear classifier's baseline (Section 3.5). *Regular* denotes the performance of the GRU in the presence of missing data using a frozen classifier trained on pristine data. *curri-MAE* denotes our MAE trained using a curriculum (Section 3.3). Finally, *Regular-ft-all*, *curri-MAE-ft* and *curri-MAE-ft-all* are the performances with a finetuned classifier (Sections 3.5 and 4.2). The absence of a baseline denotes performance below the presented range. We selected all our hyperparameters on AVID-CMA and attempted to replicate our findings on CrissCross. Of the total 12 cases we examined, curri-MAE-ft-all surpasses strong baselines in all cases except with CrissCross on RAVDESS dataset with either classifiers.

First, we focus on AVID-CMA (Figure 2). For the linear classifier (Figure 2a), we see that the finetuned classifier on MAE features (i.e., imputed) performs either favorably or equivalently to the baselines. In fact, our proposed method outperforms the baselines at all dropouts rates in CREMA-D and Lombard, most rates in RAVDESS, and is otherwise equivalent. The MAE imputation without fine-tuning, on the other hand, is mostly equivalent to *Backoff*. Interestingly, we also find that the classifier's performance improves when fine-tuned, compared to the pristine condition, which means that the MAE imputation also acted as a data *augmentation* technique. The augmentation effect is strong enough that it can persist up to high levels of dropout, such as 40% for CREMA-D and RAVDESS. We do not show the final accuracy of the classifier in the pristine condition as that is not our focus in this work.

In the GRU case, Figure 2b demonstrates the same trends.

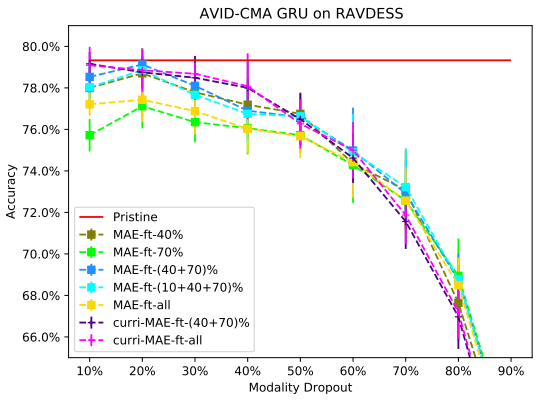


Figure 4: *Performance comparison between finetuning with curriculum MAE's (`curri-MAE`) imputations, and finetuning with fixed-rate MAE's (`MAE`) imputations, at a mixture of the designated dropout rates, i.e., `(40+70)%` denotes fine-tuning the classifier with imputations at both 40% and 70% dropout.*

Our proposed method is at least equivalent to the baselines, with CREMA-D and Lombard again showcasing the most favorable comparisons. We see fewer augmentation effects of the MAE. In the case of the GRU, it seems that the baseline fine-tuning can also occasionally provide augmentation effects. Finally, the baseline GRU with no fine-tuning degrades very rapidly.

In the case of CrissCross with a linear classifier, in Figure 3a, we see that our method continues to perform favorably in CREMA-D under all conditions, but not in Lombard and RAVDESS. Augmentation effects appear in CREMA-D.

When GRUs are used with CrissCross instead, in Figure 3b, we see that our proposed method fares favorably to the baselines, except for a single datapoint, 40% dropout in RAVDESS. Compared to AVID-CMA with a GRU, we see more noticeable augmentation effects, notably both for the MAE fine-tuning and the baseline GRU fine-tuning, in RAVDESS and Lombard. The simple GRU baseline degrades quickly except for Lombard.

4.4. Why is the Dropout Curriculum Important?

We compare a regular (70% fixed-rate) MAE with our curriculum MAE (`curri-MAE`) in Figure 4. We observe that the right dropout-rate mixture for imputing and fine-tuning the GRU can be dataset-specific with MAE. In contrast, our `curri-MAE` reliably achieves performance equivalent to the best mixture of the MAE (40%+70%), stabilizing fine-tuning. We also observe that taking MAE's best mixture and applying curriculum to it (curr-MAE-ft*) produces virtually identical results to the mixture of all rates that we use. Therefore, the proposed curriculum is integral for robustness.

5. Conclusions

In this study, we address the challenge of partial modality loss and propose a curriculum-based Masked AutoEncoder (MAE) for feature imputation. Our approach surpasses existing methods, yielding a more robust classifier. We observe that fine-tuning with imputations acts as augmentation and bridges the gap between real and generated features. As our work focuses on data sets with mostly short clips, it can be directly extended to streaming, by using a small number of future clips to derive the final embeddings of the MAE, which is left as future work.

6. References

[1] P. Morgado, N. Vasconcelos, and I. Misra, "Audio-visual instance discrimination with cross-modal agreement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 475–12 486.

[2] O. Chang, O. d. P. F. Braga, H. Liao, D. D. Serdyuk, and O. Siohan, "On robustness to missing video for audiovisual speech recognition," 2022.

[3] P. Sarkar and A. Etemad, "Self-supervised audio-visual representation learning with relaxed cross-modal synchronicity," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 8, 2023, pp. 9723–9732.

[4] Y. Tian and C. Xu, "Can audio-visual integration strengthen robustness under multimodal attacks?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5601–5611.

[5] 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.

[6] T. Srinivasan, T.-Y. Chang, L. L. P. Alva, G. Chochlakis, M. Rostami, and J. Thomason, "CLiMB: A continual learning benchmark for vision-and-language tasks," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. [Online]. Available: https://openreview.net/forum?id=FhqzyGoTSH

[7] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

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

[9] Y. Gong, A. Rouditchenko, A. H. Liu, D. Harwath, L. Karlinsky, H. Kuehne, and J. R. Glass, "Contrastive audio-visual masked autoencoder," in *The Eleventh International Conference on Learning Representations*, 2022.

[10] G. Chochlakis, G. Mahajan, S. Baruah, K. Burghardt, K. Lerman, and S. Narayanan, "Leveraging label correlations in a multi-label setting: A case study in emotion," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[11] 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*, vol. 5, no. 4, pp. 377–390, 2014.

[12] S. R. Livingstone and F. 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*, vol. 13, no. 5, p. e0196391, 2018.

[13] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. J. Brown, "A corpus of audio-visual Lombard speech with frontal and profile views," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. EL523–EL529, 06 2018. [Online]. Available: https://doi.org/10.1121/1.5042758

[14] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 695–12 705.

[15] Y. Huang, J. Lin, C. Zhou, H. Yang, and L. Huang, "Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably)," in *International Conference on Machine Learning*. PMLR, 2022, pp. 9226–9259.

[16] J. Zhao, R. Li, and Q. Jin, "Missing modality imagination network for emotion recognition with uncertain missing modalities," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 2608–2618.

[17] J. Zeng, J. Zhou, and T. Liu, "Mitigating inconsistencies in multimodal sentiment analysis under uncertain missing modalities," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 2924–2934.

[18] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "Moddrop: Adaptive multi-modal gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1692–1706, 2016.

[19] A. Hussen Abdelaziz, B.-J. Theobald, P. Dixon, R. Knothe, N. Apostoloff, and S. Kajareker, "Modality dropout for improved performance-driven talking faces," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, ser. ICMI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 378–386. [Online]. Available: https://doi.org/10.1145/3382507.3418840

[20] S. Parthasarathy and S. Sundaram, "Training strategies to handle missing modalities for audio-visual expression recognition," in *Companion Publication of the 2020 International Conference on Multimodal Interaction*, ser. ICMI '20 Companion. New York, NY, USA: Association for Computing Machinery, 2021, p. 400–404. [Online]. Available: https://doi.org/10.1145/3395035.3425202

[21] C. Ma, S. Tschiatschek, K. Palla, J. M. Hernandez-Lobato, S. Nowozin, and C. Zhang, "Eddi: Efficient dynamic discovery of high-value information with partial vae," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4234–4243.

[22] A. Nazábal, P. M. Olmos, Z. Ghahramani, and I. Valera, "Handling incomplete heterogeneous data using vaes," *Pattern Recognition*, vol. 107, p. 107501, 2020.

[23] A. Haliassos, P. Ma, R. Mira, S. Petridis, and M. Pantic, "Jointly learning visual and auditory speech representations from raw data," in *The Eleventh International Conference on Learning Representations*, 2022.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[25] H. Zheng, Z. Lin, J. Lu, S. Cohen, J. Zhang, N. Xu, and J. Luo, "Semantic layout manipulation with high-resolution sparse attention," *arXiv preprint arXiv:2012.07288*, 2020.

[26] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[27] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211.

[28] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *ICML*, vol. 2, no. 3, 2021, p. 4.

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

[30] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[31] 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)*. IEEE, 2017, pp. 776–780.

[32] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[33] S. A. Zollinger and H. Brumm, "The lombard effect," *Current Biology*, vol. 21, no. 16, pp. R614–R615, 2011.