

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

Learning Self-Supervised Representations of Audiovisual Human-Centric Data

Author: Abhinav Shukla

1st Supervisor: Maja Pantic

2nd Supervisor: Stavros Petridis

March 2023

Thesis submitted in part fulfilment of the requirements for the degree of PhD in
Computing and the Diploma of Imperial College London.

Declaration of originality

I certify that the contents of this thesis are entirely my own work, and unless otherwise indicated by appropriate references, describe my own research. This thesis has not been submitted for any other purposes than for the PhD Degree at Imperial College London

Abhinav Shukla

March 2023

Copyright declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

Deep learning models have driven significant progress in audio and audiovisual machine learning but rely on extensive supervision from large labeled datasets. While labels are often intractable to obtain, unlabeled data is significantly more pervasive and ever growing. Self-supervised learning leverages the implicit structure of unlabeled data to learn better representations that generalize better to smaller labeled datasets of interest. This is relevant in domains like emotion and speech recognition with limited labeled data, where training from scratch is infeasible and exhaustive supervision is unavailable.

In this thesis, we study the effectiveness of audiovisual self-supervised learning for learning better unimodal audio representations. The core idea is to leverage the natural synchrony and cross-modal correspondence between audio and video as a source of cross-modal self-supervision. Audio-only self-supervised learning works well, but complementary addition of visual self-supervision produces better representations. We study speech-driven facial reconstruction as a pretext task for audio representation learning in speech and find that the correlation between lip movements in video and speech content in audio; and that between facial expressions in video and speech emotion in audio provides a very rich supervisory signal. We propose new methods for audiovisual self-supervised learning for more general audiovisual content that are based on combining pretext tasks like clustering, invariance-based contrastive learning, and audiovisual sound source localization. We demonstrate strong results on a number of audio recognition tasks in both audiovisual speech data and general audiovisual content. Compared to existing methods, our proposed representations achieve better performance with smaller amounts of pretraining data while getting even better with additional data; are more robust in noisy conditions; and perform better on downstream tasks with limited labeled data. The work provides compelling evidence for the utility of self-supervision in general, and visually guided self-supervision in particular, for learning better representations of audio.

Acknowledgments

None of my growth or development as a student, researcher, or thinker would have been possible without the guidance of my mentors and teachers. I am deeply grateful for everything that they have taught me, and they continue to inspire me to want to contribute to science and the creation of new knowledge.

First and foremost, I would like to thank my PhD advisors: Maja Pantic and Stavros Petridis. Their support, guidance, patience and kindness even in difficult times was very valuable in the last few years for me. Maja gave me the opportunity to pursue doctoral research fresh out of undergraduate education when I was spirited but naive in many ways. Her advice and motivation has taught me many things about research and the world, including setting higher standards for my own work, being able to gauge the most important directions to look into, and being strategic on what to work on in a rapidly moving field. Her mentorship and support has been invaluable. Stavros gave me very deep technical mentorship while introducing me to important research areas. He inspires me by his creativity and intuition for the best ideas, and has been an exemplary role model for non-technical aspects like collaboration and the right kind of communication. Thank you both for being my advisors and for your confidence in me, teaching me, and supporting me in many different ways.

A year after I began my PhD, the world fundamentally changed as we knew it with the COVID pandemic which created an unprecedented set of challenges. Despite this, I was fortunate to have friendship, support and mentorship from my labmates and other peers, which was an important part of keeping my research interests alive. I thank my Imperial labmates Konstantinos Vougioukas, Rodrigo Mira, Pingchuan Ma, Linh Tran, Markos Georgopoulos, Eftychia Fotiadou, and Akis Kefalas. They have all contributed to this journey in their own important ways. Teresa Ng and Amani El-Kholy for making several iBUG and DoC threads seamless. To everyone from CSG for making life a lot easier.

I have been very fortunate to have so many great mentors who have made these last few years of learning a lot better. I am grateful to Anurag Kumar for believing in me and giving me a very important opportunity. I have learned so much from him about so many different things. His mentorship, guidance, and unwavering personal support continue to be invaluable. Vamsi Krishna Ithapu for supporting me and making me grow both professionally and personally. Hao Jiang and Buye Xu for their exemplary quality of work and strong technical mentorship. Fiona Ryan for being a very fun and immensely capable teammate to work with.

My time at IIIT Hyderabad was when and where I decided that I liked doing research. I am tremendously grateful to Ramanathan Subramanian for being the person who introduced me to doing academic research and teaching me so much as an undergraduate. He has been crucial to developing my interest in research. Mohan Kankanhalli for a memorable internship and valuable lessons that continue to make even greater sense with the passage of time. Maneesh Bilalpur for being there in Singapore. Harish Katti for being a mentor. CV Jawahar for some important feedback and inspiration. Vineet Gandhi for introducing excitement about working with images and visual data. Abhinav Dhall for an important conversation that helped me decide to apply for a PhD.

My friends have provided me with surroundings that have made my life so much better. Thank you Dhruva for being an amazing friend, teammate, and everything else. You are the closest thing to a real brother that I have. Sangeetha for taking so many random classes with me and always being there for everything. Shraddhan, Moneish, Arpan, Jalpreet for their constant friendship and support. To everyone else in Wing3G for their companionship and banter. Saurabh, Parth, Rohan, and everyone else of my football juniors and teammates for being part of some of my favorite memories. Anurag Ghosh and Siddhartha Gairola for sharing this research journey and being phenomenal friends who inspire me in so many ways. I can't wait to work with and learn from you in the years to come. Sanjana for her friendship and support in Hyderabad and Seattle. Roopal for being there and motivating me to actually write this thesis. Aditya, Garima, Shreya, Karthik, Sneha, Amala, Sheshansh, Soumya, Gantavya, and Sahil for being Seattle friends. Shruti for having been a very important part of my life, and teaching me so much about so many things while growing with me as an individual, being patient and supportive. Sukriti for her love, unwavering support, and being my biggest cheerleader through thick and thin. I am incredibly thankful to have met you and can't wait to learn, grow, and make memories together. Thank you for making everything better :)

I would be nothing without the love and support from my family. To my parents and sister, you have gone above and beyond to be my biggest supporters and cornerstones. You have always allowed me to pursue my evolving passions and been there for everything. To everyone in Sevagram who supported me through childhood and more, thank you for the wonderful environment and memories!

To my family for always being there for me

Contents

1	Introduction	1
1.1	Self-Supervised Learning: What and Why?	3
1.2	Why Self-Supervised Learning for Audiovisual data in particular? .	5
1.3	Key Contributions and Ideas	7
1.4	Thesis Organization	8
2	Background	10
2.1	Multimodal Machine Learning	10
2.1.1	Audiovisual Speech Data Format	11
2.1.2	Audiovisual Speech Recognition	11
2.1.3	Audiovisual Emotion Recognition	12
2.1.4	Speech-driven Facial Reconstruction	13
2.2	Self-Supervised Learning: The Concept	13
2.2.1	History	13
2.3	Unimodal Self-Supervised Learning	15
2.3.1	Self-Supervised Learning for Visual Data	15
2.3.2	Self-Supervised Learning for Text Data	18
2.3.3	Self-Supervised Learning for Audio and Speech Data	18
2.4	Multimodal Self-Supervised Learning	21
2.4.1	Audiovisual Self-Supervised Learning	21
2.4.2	General Cross-Modal Self-Supervision	24
2.4.3	When does Multimodal Self-Supervision not work?	24
2.5	Applications of Self-Supervised Learning for Audio and Speech . .	25
2.5.1	Speech Recognition and Enhancement	25
2.5.2	Affect/Emotion Recognition	25
2.5.3	Audio Event Detection	26
2.5.4	Other Audio Applications	26

3	Audio-only Self-Supervised Learning of Speech Representations	27
3.1	Background	28
3.2	Audio Self-Supervision for Speech Representation Learning	30
3.2.1	Arrow of Time for Audio	30
3.2.2	Odd One Out for Audio	31
3.2.3	Reconstructing Salient Speech Attributes	31
3.2.4	Audio Encoder Architecture	31
3.3	Experiments and Results	32
3.3.1	Datasets	32
3.3.2	Baselines	34
3.3.3	Experimental Setup	37
3.3.4	Results	38
3.4	Discussion	39
4	Visually Guided Self-Supervised Learning of Speech Representations	40
4.1	Background	40
4.2	Speech-Driven Facial Video Reconstruction for Self-Supervision . .	41
4.2.1	Why Facial Reconstruction is Relevant for Emotion and ASR	42
4.3	Experimental Setup	43
4.3.1	Datasets	43
4.3.2	Baselines	44
4.3.3	Evaluation	45
4.4	Results	46
4.5	Discussion	47
5	Joint Audiovisual Self-Supervision for Speech Representation Learning from Log Mel Inputs	50
5.1	Background	52
5.2	Combining Audio and Visual Self-Supervision	52
5.2.1	Visual Self-Supervision for Speech Representation Learning	52
5.2.2	Why Visual Self-Supervision for Emotion Features?	54
5.2.3	Audio Encoder Architecture	56
5.2.4	Audiovisual Self-Supervision for Speech Representations . .	56
5.3	Experimental Setup	57
5.3.1	Datasets	57
5.3.2	Baselines	58
5.4	Results	59

5.4.1	Experimental Setup	59
5.4.2	Results with Visual Self-Supervision (L1)	64
5.4.3	Results with Audio-only Self-Supervision (Odd)	65
5.4.4	Results with Audiovisual Self-Supervision (L1 + Odd)	65
5.4.5	When does Visual Self-Supervision help specifically?	66
5.4.6	Performance in various levels of noise	72
5.4.7	Performance with various sizes of the pretraining set and downstream training set	72
5.4.8	Comparison with finetuned and supervised versions	74
5.5	Discussion	75
5.5.1	Limitations	76
5.5.2	Future work	76
6	Joint Audiovisual Self-Supervision for Speech Representation Learning from Raw Audio	79
6.1	Background	79
6.2	Method	81
6.2.1	Audio Encoder Architecture	81
6.2.2	Visual Self-Supervision	81
6.2.3	Audio Self-Supervision	82
6.2.4	Audiovisual Self-Supervision	83
6.3	Experiments	84
6.4	Discussion	86
6.5	Audio Encoders	87
6.6	Dataset and split details	89
7	Audiovisual Self-Supervised Learning of General Sound Representa- tions	91
7.1	Background	92
7.1.1	Related work in Visual and Audiovisual Self-Supervision	93
7.2	Self-Supervised Learning for Generic Audiovisual Data	98
7.2.1	AVCC - Pseudolabel based clustering with Memory Banks	98
7.2.2	Audiovisual SimSiam	101
7.2.3	Audiovisual Barlow Twins	102
7.3	Experiments and Results	105
7.3.1	Datasets	105
7.3.2	Ablation Studies	106

7.3.3	Downstream Audio Results	110
7.4	Discussion	114
8	Conclusion	116
8.1	Contributions	116
8.2	Limitations	118
8.3	Possible Future Directions	119

List of Tables

3.1	Results for audio-only self-supervised methods on discrete emotion recognition on CREMA, Ravdess, IEMOCAP	34
3.2	Results for audio-only self-supervised methods on continuous affect recognition on SEWA and RECOLA	35
3.3	Results for audio-only self-supervised methods on speech recognition on SPC	36
4.1	Pretraining and evaluation dataset details for Chapter 4	43
4.2	Results for discrete emotion recognition on CREMA and Ravdess for audio features guided by visual self-supervision	46
4.3	Results for speech recognition on GRID and SPC for audio features guided by visual self-supervision	47
5.1	Pretraining and evaluation dataset details for Chapter 5	59
5.2	Results for discrete emotion recognition on CREMA, Ravdess, and IEMOCAP for audio features guided by audiovisual self-supervision	60
5.3	Results for continuous affect recognition on SEWA and RECOLA for audio features guided by joint audiovisual self-supervision	61
5.4	Results for speech recognition on SPC for audio features guided by joint audiovisual self-supervision	62
5.5	Relative CCC gains achieved by audio, visual, and join audiovisual self-supervision for high and low valence and arousal sequences on RECOLA	68
5.6	Ablation study for weight tuning for the multi-task audiovisual loss function	70
5.7	Results for speech recognition with limited downstream labeled data on SPC with joint audiovisual self-supervision	74

6.1	Results for speech recognition on SPC and LRW for audio features guided by joint audiovisual self-supervision from raw audio inputs	82
6.2	Results for speech recognition on SPC and LRW under varying levels of introduced babble noise	85
6.3	Encoder type and number of trainable parameters in each of the compared methods in chapter 6	87
6.4	Feature dimensionality and audio sample rate of each method compared in chapter 6	88
6.5	Pretraining dataset details for each method compared in chapter 6	89
6.6	Evaluation dataset splits for chapter 6	89
6.7	Evaluation dataset split durations for chapter 6	89
6.8	Evaluation dataset class statistics for chapter 6	90
7.1	Ablation study for optimal value of number of clusters in AVCC . .	107
7.2	Ablation study for sizes and types of pretraining datasets for AVCC	109
7.3	Ablation study for feature dimensionality used for the Barlow Twins-based methods	110
7.4	Downstream results on environmental sound classification on ESC50 and speaker identification on Voxceleb1	111
7.5	Downstream results for audio event detection on the balanced Audioset	113

List of Figures

2.1	An illustration of the idea behind RotNet by Gidaris et. al. [70] . . .	17
2.2	An illustration of the concept of colorization as a free supervisory signal from Vondrick et al.[205]	17
2.3	An illustration of Shuffle and Learn from Misra et el.[130]	18
2.4	An illustration of PASE by Pascual et al.[146]	20
3.1	An illustration of Odd One Out networks from Fernando et al.[63]	29
3.2	An illustration of our proposed variation of Odd One Out networks for audio	32
4.1	An overview of our proposed model for learning audio representations using speech-driven facial reconstruction	43
4.2	Detailed view of our proposed network for audio representation learning using speech-driven facial reconstruction	49
5.1	An overview of our proposed model for visually guided self-supervised audio representation learning from log mel inputs	51
5.2	A detailed illustration of our the encoder-decoder model we use for facial video reconstruction from log mel inputs	53
5.3	Frames from videos produced by the proposed method driven by the same sentence spoken in different emotions	55
5.4	Classwise performance gain of audiovisual vs. audio-only self-supervision for discrete emotion recognition	67
5.5	Histogram of classwise performance gain for audiovisual vs. audio-only self-supervision for speech recognition	69
5.6	Results for speech and emotion recognition under various levels of noise	71
5.7	Ablation study for speech and emotion recognition with different sizes of the pretraining set	78

6.1	An illustration of the encoder-decoder model we use for joint audiovisual self-supervision from raw audio	80
6.2	A detailed illustration of the encoder-decoder model we use for lip video reconstruction from raw audio	87
6.3	A detailed illustration of the encoder-decoder model we use for audio-only self-supervised representation learning from raw audio .	88
7.1	A detailed explanation of MoCo (Momentum Contrast) method by He et al. [81, 82]	94
7.2	An illustration of the work GDT by Patrick et al. [148]	95
7.3	An illustration of the work on Cross-Modal Deep Clustering (XDC) by Alwassel et al. [8]	96
7.4	An illustration of our proposed method AVCC (AV Contrastive + Clustering) which combines the concepts of GDT [148] and XDC [8] by introducing additional positive samples into the memory banks by clustering	99
7.5	A detailed illustration of our proposed method AVCC (AV Contrastive + Clustering)	100
7.6	An illustration of our proposed variations of the SimSiam [36] method to learn audio and audiovisual representations	101
7.7	The three separate variations of our proposed methods that build on the original Barlow Twins work [222] for audio and audiovisual representation learning combined with audiovisual sound source localization	103

Chapter 1

Introduction

The 2010's will be remembered in computer science as the decade in which supervised deep learning took the technology industry by storm. Holistic progress in powerful hardware accelerators, distributed systems, the availability of an enormous amount of new multimedia data over the internet, and the implementation of scalable machine learning systems building on existing machine learning theory led to a perfect storm which made a lot of supervised deep learning applications tractable. The AlexNet [105] work was the first mainstream deep neural network trained on the ImageNet [52] dataset which offered compelling evidence that the performance gains attained by using large-scale deep learning methods (that were conceptually simpler and more general compared to the existing specific hand-crafted state-of-the-art machine learning methods) were significant enough to warrant a complete paradigm shift in how we thought about machine intelligence. The ImageNet moment for visual recognition led to a cascading and significant effect across other disciplines such as natural language processing (with 2017 being the year the Transformer [204] architecture became prevalent), audio processing (where deep convolutional networks like the ResNet [80] operating on spectrograms treated as images became the go-to method), reinforcement learning, and multimodal machine learning. Basically any field which was based on any sort of machine learning model anywhere in its processing pipeline saw significant performance gains. Deep learning had arrived and was here to stay. The elephant in the room however, was the extremely strong dependence on large scale and high quality labeled datasets that were necessary to optimally train these models. The curation of these datasets became a critical bottleneck, but was necessary to facilitate the training of deep neural networks that generalize to similar recognition tasks.

Techniques like transfer learning (either by frozen feature extraction followed by learning simpler downstream classifiers, or directly finetuning on smaller downstream datasets) helped domains that did not have access to large scale data or those in which data was prohibitively expensive or required significant domain expertise to label. But the core reliance on large scale labeled data remained a persistent issue, and labeled data for a lot of domains was very hard to come by (especially for things like speech recognition for limited labeled data and audiovisual emotion recognition; which are two downstream tasks that this thesis focuses on).

At the same time, there has been an exponential and explosive growth in the amount of *unlabeled* multimedia data being produced and collected. Whether it be audiovisual content on video sharing and streaming platforms, petabytes of text being added to online forums and social media websites, thousands of hours of speech data being recorded in call centres, or even time-series data from sensors being recorded in a healthcare setting. The collection and storage of unlabeled multimodal data is both voluminous and pervasive at an unprecedented scale. It is simply not possible to manually label a large portion of this data (for a variety of reasons) to use for training deep neural networks in a supervised way.

However, this unlabeled data is still inherently structured and has intrinsic patterns in its distribution which still offer a valuable way to process it by deep neural networks which are great at discovering and exploiting these patterns. Even classical unsupervised methods like Principal Component Analysis (PCA) offer insights into smaller-scale datasets. But the challenge now was to marry the merits of the deep learning paradigm shift with the lack of available manually annotated datasets while still being able to effectively represent the data of interest. So, in the 2020's, with the current landscape of machine learning, compute, and data; how do we effectively learn useful representations of large-scale unlabeled datasets in a way that makes them generalize for smaller and more specific labeled datasets?

A promising avenue is **Self-Supervised Learning** (SSL). The ideas behind SSL are not new, with some works as early as the 1980's laying down the theoretical foundations (see Section 2.2). Being able to scale up these simple ideas well however, has led to remarkable advances in representation learning (simple in algorithmic terms like using contrastive learning, or a proxy cross-entropy loss).

1.1 Self-Supervised Learning: What and Why?

So what really is ‘*self-supervised learning*’? This is a fairly contentious question in the machine learning community with varying opinions (e.g. some academics think that self-supervision is not strictly unsupervised learning and is a variant of supervised learning). In my opinion: **self-supervised learning is a form of unsupervised learning in which the encoder model is trained by a proxy supervised learning problem**. This holds true for a wide variety of contemporary self-supervised learning methods, both for invariance-based (e.g. contrastive learning for two separate augmentations of the same sample) methods and for generative (e.g. masked language modeling) methods [14]. The key qualifier about a method being self-supervised is that it does not in any way use manually annotated human labels in any step of the learning process. The labels used to provide the ‘proxy supervision’ come from the data itself (e.g. reconstructing the masked portion of a sequence, bringing the representations of two different augmentations of the same data point closer together, assigning pseudolabels by clustering the data followed by predicting the cluster IDs, and many other ways). The form of supervision provided by a typical supervised learning problem is by the categorical label of the data, but for self-supervised learning it is the *inherent structure* of the data which can provide a richer signal to learn the underlying distribution of the data.

The core ideas behind self-supervised learning (or more generally unsupervised learning) are not very new. These ideas (as with most other contemporary machine learning ideas) have their genesis in the past and have only recently come to fruition due to the inflection of various necessary developments in other fields (hardware, data, distributed systems) which have also fueled the supervised learning revolution (related work including these older methods is explored in depth in Chapter 2).

Why do we not just use large labeled datasets for the rest of time? The introduction section of one of my all-time favorite doctoral theses [128] contains a fantastic explanation for this question. Labeling for the scale of multimedia data that we produce and consume, is extremely expensive in the best cases, and downright impossible in most cases (e.g. over 3 Billion people use apps like Facebook, Instagram, YouTube, Twitter daily and produce a gigantic and never ending stream of new data). There is a critical need to be able to perform general-purpose representation learning for this data without involving extensive manual annotation.

There are also some specific considerations for why learning from unlabeled data may be useful even if labeled datasets are available (discussed in [128]):

- **Labels are not always enough:** Labeled datasets are often curated with a very specific ontology. The choice of this may artificially impose a certain upper limit of how good the representations learned by just supervised prediction of these classes can end up being. The actual number of concepts/categories/labels that may be applicable to a particular modality is often far richer and more diverse than the ones found in the large scaled datasets. For instance, the ImageNet-1k dataset [52] has 1000 classes with a lot of the classes being specific animal/plant categories which may not be of interest for learning representations for other domains. The MS-COCO dataset [111] on the other hand contains 91 classes of more common objects with concepts that are more generally applicable to human scenarios. However, depending on the specific machine learning task of interest, these labels may still not be representative enough. There are millions of visual or multimodal concepts (e.g. hashtags on Instagram, compositional semantic image captions) that are all slightly different and might not be accurately represented in current large scale labeled datasets.
- **Samples from the long tail:** There are a lot of data instances that come from the long tail parts of the data distribution. These instances are particularly challenging to curate and label, while also often being important to represent well. These occurrences (for which we cannot realistically expect accurate or high quantity labels) are far more likely to be seen in larger unlabeled datasets than in labeled datasets. Self-supervision is better equipped to deal with this scenario than supervised deep learning.
- **Artificially balanced datasets:** Balanced datasets are much easier to train deep learning models on, especially in a supervised setting. Large datasets for image recognition (e.g. ImageNet [52]), video understanding (e.g. Kinetics [95]), audiovisual speech recognition (e.g. LRW [41]), and datasets for many other problems have the curious issue of having been artificially balanced. This makes the training of supervised models on these tasks much easier, but could be potentially detrimental to the learning of the best representations on these datasets because the true distribution is much more imbalanced.

Self-supervised learning directly helps the large-scale learning of representations from massive unlabeled datasets, without the need of manual supervised labels / human annotations. This is especially relevant in the context of domains

that have a limited amount of labeled data. There are a large number of such domains, but of particular interest to this thesis are speech and audio (visual and text self-supervised learning was a lot more common than audio at the time of conducting research for this work), especially from the lens of downstream problems like emotion recognition and speech recognition for limited labeled data (both of which closely fit the archetype of a problem that can see huge gains by self-supervision). Both of these domains are plagued by the lack of availability of high quality and large labeled datasets.

This thesis focuses on learning self-supervised representations of audiovisual data. A specific direction of inquiry that we undertake is to learn unimodal representations of audio for audio understanding tasks by using multimodal self-supervision in which the video modality provides auxiliary self-supervision and guides the representation learning process. Unlabeled audiovisual data (or more generally paired multimodal data) is especially compelling for self-supervised representation learning due to a number of reasons that we will discuss now.

1.2 Why Self-Supervised Learning for Audiovisual data in particular?

Audiovisual data (typically a video clip with paired audio and visual data) is an extremely good candidate for a variety of self-supervised learning methods. It has a few unique advantages over unimodal data such as images, audio, and text. **The natural synchrony between the two modalities combined with the complementary information provides a large amount of free cross-modal supervision.** Both invariance-based and generative methods can leverage this cross-modal correspondence for self-supervised learning in ways that are not possible for unimodal data. Let us dive deeper into these aspects:

- **Natural synchrony:** The two modalities have natural synchrony and co-occurrence of salient events. The fact that these sound events happen at the *same timestep in both modalities* is vital. For example in audiovisual speech, lip movements in the visual modality are correlated with speech content in the audio modality. In sports videos, the sound of a ball bouncing in the audio modality comes with the actual visual of the ball bouncing in the video modality. In egocentric task videos like cooking, the clink sound of a spoon hitting a utensil is accompanied with the visual of the same thing in the

video. All of these are examples of this natural cross-modal synchrony which can provide a very powerful learning signal to be used by self-supervised learning. Merely having access to a large amount of data like this provides a huge source of proxy labels (e.g. one modality can be predicted from another to learn better representations for the first modality). Creatively leveraging this natural synchrony in audiovisual human-centric data (audiovisual speech in particular) forms the crux of this thesis.

- **Complementary information:** While the two modalities contain useful information on their own, the type of information is often complementary. A classic example of this comes from audiovisual speech and the McGurk effect [125]. From a human perception standpoint, having access to both modalities improves speech comprehension when compared to using either modality individually. Access to multimodal information also helps greatly when one modality is influenced by noise (e.g. for speech recognition, having access to lip movements visually can help in cases where the audio is noisy). More generally, whenever we have access to multiple complementary modalities and are able to correctly model them, we are able to get better downstream task performance than when using any individual modality.
- **Ability to use redundancy and complementarity together:** Self-supervised learning on audiovisual data benefits from being able to leverage the natural cross-modal redundancy (e.g. by predicting one modality from another or by making the representations for the two modalities come closer in the latent space). However, it also benefits from being able to use the extra complementary information available in each modality as a source of auxiliary supervision (e.g. using lip movements to learn better representations of audio even when the audio is noisy). This setup is unique to audiovisual data specifically and paired multimodal data in general. However, audiovisual data contains extremely fine grained cross-modal synchrony that is not as commonly observed in other mainstream multimodal self-supervised setups (e.g. in vision + text self-supervised learning). Creatively using both the synchronicity and the complementarity is a unique advantage with audiovisual data and self-supervised learning for audiovisual data (and especially audiovisual speech) has only recently been begun to be explored deeply.

This thesis focuses on self-supervised learning for audiovisual human-centric data in general, and audiovisual speech data in particular. The core idea is to

enable the learning of better audio representations by using the visual modality to guide the representation learning process. Doing this well has important applications in downstream audio tasks such as speech recognition, speech emotion / affect recognition, speaker identification, audio event detection, and potentially any other machine learning task which is based on representations of audio that may have certain settings in which paired unlabeled video data is available.

1.3 Key Contributions and Ideas

This thesis studies self-supervised learning for audiovisual human-centric data, and audiovisual speech in particular. The work done for this thesis makes the following major research contributions:

1. **Visually guided self-supervised learning of audio features:** The most important contribution of this thesis is to show that speech-driven facial reconstruction can be effectively used for visually guided self-supervised learning of audio features in audiovisual speech data. The method is particularly relevant for downstream speech recognition and speech emotion recognition, because it is able to leverage the **correlation between lip movements and speech content, and between facial expressions and speech emotion**. The audio features are *guided* by the visual modality during the pretraining process, and achieve excellent performance on downstream audio tasks (speech recognition, speech emotion / affect recognition). The proposed audio features outperform training from scratch (supervised deep learning) and even outperform finetuning after supervised pretraining. This contribution was presented in the ICASSP 2020 paper (Shukla, Vougioukas, Ma, Petridis, and Pantic) [187] and expanded in the Transactions on Affective Computing 2021 paper (Shukla, Petridis, and Pantic) [188].
2. **Joint audiovisual self-supervision to learn audio features for speech:** This work demonstrates that a combination of audio and visual self-supervision in audiovisual speech can learn better representations compared to unimodal audio-only or visual-only self-supervision. The audio representations learned by using a combination of audio and visual self-supervision not only outperform both training from scratch and supervised pretraining in clean conditions, but are also more robust to noisy conditions for both speech and emotion recognition across all levels of noise, and for all sizes of pretraining data. The proposed representations are also significantly better than

the baselines in all tested settings when using a small fraction of labels in the downstream datasets. This contribution was presented in the ICML SAS 2020 paper (Shukla, Petridis, and Pantic) [187] for raw audio and the Transactions on Affective Computing 2021 paper (Shukla, Petridis, and Pantic) [188] for mel features, with a shorter version appearing in CVPR 2020 Sight and Sound (Shukla, Petridis, and Pantic) [185].

3. **Variations of contrastive self-supervised learning for audio and audiovisual data:** Although the core of this thesis is based on a generative cross-modal self-supervised learning method, we also demonstrate how recently proposed state-of-the-art methods for unimodal contrastive self-supervised learning can be adapted to a multimodal and cross-modal representation learning setup. The proposed contrastive methods that use a combination of audio-only and audiovisual self-supervision outperform their unimodal counterparts. This contribution was demonstrated in Chapter 7 and was conducted during an internship at Meta Reality Labs Research.
4. **Audiovisual sound source localization for self-supervision:** Finally, this work also explores the utility of self-supervised sound source localization for audio representation learning. Just as the lip regions and facial expressions in the video modality specifically provide visual self-supervision for audio representation learning in speech data, the sound producing regions of the video in datasets like AudioSet provide auxiliary self-supervision for more generic audiovisual data. This leads to performance gains on problems like audio event detection and speaker identification. This contribution was demonstrated in Chapter 7 and was conducted during an internship at Meta Reality Labs Research.

1.4 Thesis Organization

The thesis is organized in 8 total chapters (including this introduction chapter):

- **Chapter 2** introduces the relevant background and historical context in the areas of multimodal machine learning, self-supervised learning, and specific applications of these in the audio and speech domains.
- **Chapter 3** dives deeper into unimodal audio-only self-supervised learning of speech representations, while proposing two new methods based on temporal order verification.

- **Chapter 4** introduces the crux of this thesis, which is using the visual modality as a supervisory modality to guide the learning of audio representations, especially when speech-driven facial reconstruction is used as a pretext task.
- **Chapter 5** develops on the ideas introduced in previous chapters to perform joint audiovisual self-supervised learning (using a combination of audio-only, video-only, and cross-modal losses) which helps learn a better overall speech representation from log mel inputs.
- **Chapter 6** explores joint audiovisual self-supervised learning directly from the raw audio waveform and by using the prediction of salient audio attributes as the audio pretext task.
- **Chapter 7** explores modern self-supervised contrastive learning methods for multimodal representation learning of general sound representations (as opposed to only speech-representations). It also introduces multiple proposed methods, while also studying audiovisual sound source localization as a potential self-supervised learning objective.
- **Chapter 8** summarizes the key contributions and takeaways of the thesis.

The relevant publications associated with the thesis are:

1. Visually Guided Self Supervised Learning of Speech Representations [187] (ICASSP 2020).
2. Visual Self-Supervision by Facial Reconstruction for Speech Representation Learning [185] (Sight and Sound Workshop, CVPR 2020)
3. Learning Speech Representations from Raw Audio by Joint Audiovisual Self-Supervision [186] (Workshop on Self-supervision in Audio and Speech, ICML 2020)
4. Learning Self-Supervised Multimodal Representations of Human Behaviour [180] (ACMMM 2020 DC)
5. Does Visual Self-Supervision Improve Learning of Speech Representations for Emotion Recognition? [188] (IEEE Transactions on Affective Computing 2021)

Chapter 2

Background

2.1 Multimodal Machine Learning

A seminal and detailed introduction of Multimodal Machine Learning can be found in the taxonomy work by Baltrusaitis et al. [18]. This work establishes 5 key challenges in multimodal machine learning: representation, alignment, translation, fusion, and co-learning. Out of the discussed application areas of multimodal machine learning, representation was the most pervasive as it was a necessary step for every single problem. In this thesis, we want to directly focus on learning and evaluating better representations of multimodal (audiovisual) data. Of particular interest to us is audiovisual speech data which is a very crucial type of data to represent and understand human behaviour.

Audiovisual speech data (especially the unlabeled variety) is extremely common and the usage of complementary information from both modalities is a popular concept in many fields of research. The McGurk effect [125] was the classic example that demonstrated the audio-visual nature of human perception of speech. The visual modality contains information that offers robustness in circumstances where the audio modality may be corrupted with noise [152]. In this thesis, we focus primarily on developing better methods for learning audio representations in audiovisual speech data that are guided by the visual modality. To better position the contributions of our work, we will now introduce literature in the subareas of multimodal machine learning that our work is very relevant to: audiovisual speech recognition; audiovisual emotion recognition; and speech-driven facial reconstruction.

2.1.1 Audiovisual Speech Data Format

In this thesis, “audiovisual speech” data refers to a time-aligned sequence of audio and video samples of frontal-facing cropped faces. Unless otherwise specified, the sampling rate is 25 FPS for the video modality, and 16 kHz for the audio modality. The visual fidelity and audio SNR are specific to the used datasets. The resolution of the cropped faces is 256 pixels high and 128 pixels wide. For all the experiments, the data was stored as separate .mp4 files for the video, and .wav files for the audio (two files per individual dataset item).

2.1.2 Audiovisual Speech Recognition

Audiovisual speech recognition is a classic example of a multimodal machine learning problem which benefits from complementary audio and visual information. In addition to the McGurk effect [125], there has been also historical work that shows that visual cues are vital in the process of language learning in infants [126] and that visual contribution to speech intelligibility in noise is significant [192].

Modern end-to-end speech recognition systems are usually based on optimizing the CTC loss function [73], which computes a loss between a continuous time series (temporal speech features) and a target sequence (target words, characters, or phonemes). Recent work has combined the CTC loss function with attention (Bahdanau attention [17]) mechanisms for hybrid audiovisual speech recognition with improved results [153]. The state of the art method for audiovisual speech recognition until very recently was based on the popular RNN-T (Recurrent neural network transducer) framework [123]. Recently, there has been a shift to using conformer based architectures which have achieved better performance on audiovisual speech recognition [118]. There has been a lot of work which has highlighted performance gains in audiovisual as opposed to audio-only speech recognition, driven by the complementary information in the video that also offers additional robustness in the presence of auditory noise [217, 152, 1].

There has also been a wave of recent work on lipreading, or visual-only speech recognition [43, 179, 151, 13, 3, 35]. Fernandez et al. present a survey on lipreading approaches [62]. Visual cues have also been utilized to improved sign language recognition [7], which shares underlying principles with audiovisual speech recognition but with gestures as the base modality instead of audio. Lipreading has benefited from cross-modal distillation [4], has been shown to work for multiple languages in the wild [119], has been shown to work at the sub-word level

[158], and on both large scale datasets like LRS2 [43] and LRS3, [2] and small-scale datasets [154].

In this work, we do not focus on the related task of lipreading which uses audio features to learn better video features for visual-only speech-recognition. Instead, we try and do the opposite problem which is to learn better audio features that can leverage the information present in the visual modality for better audio-only speech recognition.

2.1.3 Audiovisual Emotion Recognition

Audiovisual emotion recognition has also seen a significant amount of recent research efforts. Automatic affect recognition has a variety of applications in various fields; from detecting depression [46], to more emotionally relevant advertising [182, 183, 184]. A lot of contemporary affect analysis approaches are based on deep neural networks that study both the visual and audio modalities [102, 221, 181]. There is also specific work that is focused on audiovisual machine learning for subsequent audio emotion recognition for both speech and non linguistic vocalizations [150, 225]. However a big problem in emotion recognition is the lack of reliably annotated data for large datasets, which we try to address (implicitly) in this thesis.

Emotion recognition can be cast in different ways, but in this thesis we primarily focus on either discrete emotion recognition based on the Ekman theory of basic emotions [57] which casts emotions as discrete well defined and universal categories such as anger, sadness, happiness etc; or on continuous affect recognition based on Russell’s circumplex model of emotions [172] which sees emotions as a continuous gradation of valence and arousal on a 2D grid (and sometimes with additional axes such as dominance). For our experiments, we evaluate audio representations on their relative effectiveness in predicting discrete emotions and continuous valence and arousal. We are particularly interested in improving audio representations by using visual self-supervision which can utilize cross-modal correspondences between facial expressions in video and speech emotion in audio.

There are numerous modern datasets for multimodal emotion recognition (especially audiovisual) which are suited for our task, including the MOSI [220], MOSEI [221], SEWA [104], RECOLA [168], AffWild [102], CREMA [25], RAVDESS [115], and IEMOCAP [24] datasets.

Zhang et al. [224] provide a detailed overview of methods for multimodal emotion recognition and Zhu et al. provide a detailed overview of deep audiovisual

machine learning in general [228].

2.1.4 Speech-driven Facial Reconstruction

A core contribution of this thesis is to use speech-driven facial reconstruction as a novel pretext task for audio representation learning. We briefly review the history and genesis of the idea of speech-driven talking head generation. There were classic works such as “*Video rewrite: Driving visual speech with audio*” [22] in 1997 and “*Mary 101: Trainable videorealistic speech animation*” [60] in 2002 which are very early demonstrations of the idea. A more recent popular example of this is the “*Synthesizing Obama*” [194] work in 2017. The underlying idea has remained the same despite going through multiple iterations of improvements due to progress in the modeling approaches and available data for the audiovisual modalities. As the visual fidelity and realism of the artificially generated talking heads increases, there has also been recent work on being able to detect face forgery [76, 78], which is an important problem in the modern social context. Synthetic faces have also been proposed for affect analysis [101], and modality dropout has shown to be effective at increasing their affective realism [90].

Our proposed method for audio representation learning is based on recent work on speech-driven facial animation [206] and its more realistic and detailed version. [208]. There is also recent work that does the opposite problem of video driven speech reconstruction in the audio modality [207].

2.2 Self-Supervised Learning: The Concept

The most compelling definition of self-supervised learning in the modern context that I subscribe to is setting up unsupervised learning (i.e. learning without any manual labels) as a ‘proxy supervised learning’ problem, where the (pseudo) labels are provided by the data itself.

2.2.1 History

As a researcher in the 2020s working in this area, it is fascinating to see how the things that are now possible with large scale self-supervised learning, with models like DALL-E-2 [163] capturing the imagination of millions of people around the world with image generation from text prompts, audiovisual models like GDT [148] being used to recommend better Instagram reels, text encoders like BERT

[53] spearheading an entire modern natural language processing industry. Scale has truly unlocked the potential of these ideas centered around learning from unlabeled data. However, if we dive into the annals of machine learning literature, the underlying ideas have existed for a very long time! Even if the applications we see today are far removed from the era of vintage NeurIPS papers with hand-drawn tSNE-like plots, it is very interesting to see the genesis of these concepts which now form the backbone of modern unsupervised machine learning using deep neural networks.

The first mention of the term ‘*self-supervised learning*’ was in a paper on recognition of vowel sounds in 1978 [144].

Gradually around the 1990s, the term became more commonplace and found more common usage in machine learning literature. Schmidhuber [174] applies the idea to credit assignment in reinforcement learning, with the model predicting “not only the reinforcement units, but the other input units too”. This idea of predicting parts of the input data sample (either within or across modalities, using predictive or contrastive self-supervision) forms the basis of a lot of modern self-supervised learning. There has also been other historical work centered on very similar ideas (although sometimes presented differently as an enhancement to supervised learning) by Towell [199] and Goldman et al. [71]. Decoux [50] uses the idea for stereo vision correspondence in 1997.

Naturally, the ideas of self-supervision percolated to multimodal machine learning (which I believe to be one of the most relevant theoretical areas for these ideas, due to the complementary information in modalities being a perfect target for self-supervision). Yamauchi et al. [219] showed how self-supervision can be used to leverage information from multiple sensors. On the note of multimodality, to quote Virginia De Sa from the abstract of her 1994 NeurIPS paper “*Learning Classification with Unlabeled Data*” [49]:

“Unfortunately, when modeling human learning or constructing classifiers for autonomous robots, supervisory labels are often not available or too expensive. In this paper we show that we can substitute for the labels by making use of structure between the pattern distributions to different sensory modalities.”

This quote is especially striking to me because nearly three decades later, the philosophical premise of my PhD thesis is almost exactly the same. The core techniques used by most of these works (e.g. by predicting one modality from another, predicting a different part of the input data, learning a distribution over possible

next parts of the data etc.) are all pervasive in modern self-supervised learning (with the key difference being scale and computation).

While there has been (and continues to be) some degree of disagreement about the precise definition of self-supervised learning, my own understanding of the term is “setting up unsupervised learning as a ‘proxy supervised learning’ problem where the labels are implicitly provided by the data”. The key aspect here being the complete lack of use of any labels provided by human annotators.

LeCun [108] introduces a very detailed position paper on the path to autonomous machine intelligence, which is heavily based on self-supervised learning, especially in an energy based formulation.

2.3 Unimodal Self-Supervised Learning

2.3.1 Self-Supervised Learning for Visual Data

In the last few years, the visual modality has seen a surge of work in self-supervised representation learning, particularly for static images.

Gidaris et al. [70] predict rotations for unlabeled images that have been rotated by a known amount, which drives the features to encode information about the object shape and appearance. Other works try to predict the relative location of patches [55], temporal order of frames in a video [63], or use contrastive learning [28, 81, 129]. BigBiGAN is a recent method proposed for adversarial self-supervised representation learning [56]. The work shows that more accurate and realistic reconstructions tend to produce better visual features for downstream tasks. Donahue et al. [56] have also demonstrated adversarial learning as a promising avenue to learn better visual features. Cycle consistency is also a concept that has been explored for visual feature learning [211]. DeepCluster [26, 27] was an interesting idea which focused on clustering in the latent space based on iteratively improving labels provided by the model being trained. NoisyStudent [216] was a follow up work on a similar concept, the idea being that the predictions of the model from a previous training epoch could be used as labels for the current epoch. The ability to train deeper networks (due to a multiple series of advancements like BatchNorm [91] among others) and recently proposed architectures like EfficientNets [196] when combined with self-supervised learning are able to achieve state-of-the-art performance on datasets like ImageNet [52]. S4L (Self-Supervised Semi-Supervised Learning) is another recent work which combines self-supervised learning with a small amount of labeled data to learn richer

representations [223]. Contrastive learning is a recent trend in self-supervised learning that is focused on separating representations of positive and negative pairs in the latent space. MoCo [81, 82, 38] is an important work in this area, and is based on distancing a positive pair from a large memory bank of negative examples. PIRL [129] extends this idea to produce image representations that are invariant of the chosen pretext task. SWaV [28] works by contrasting pseudolabel cluster assignments. BYOL [74] and SimCLR [36] are related methods for contrastive learning from two different augmented views of the same image. SimSiam [37] further simplifies this formulation. Barlow Twins introduces a new paradigm for self-supervised learning which is based on making the cross-correlation matrix of the output features from two augmented views as close to the identity matrix as possible [222]. More recently, the transformer architecture and masked autoencoders have provided even more performance gains in this area. Caron et al. [29], Wei et al. [214] and He et al. [83] show several strong results using vision transformers and self-supervised learning. All methods discussed so far are either a form of generative or invariance-based learning. Assran et al. show how both generative and invariance-based self-supervised learning can be combined in a single architecture [14].

In addition to still images, representation learning for videos has also made great strides recently. The added temporal dimension enables a further dimension of possible self-supervision. The sheer volume of video data being constantly added and consumed on the internet also makes it a very important modality to learn representations for. Jing et al. [94] extend the idea of RotNet [70] for videos. Predicting other attributes such as temporal order [130, 215], optical flow [193], temporal cycle consistency [211], colorization [205] and using visual tracking [210] are all good supervisory signals for self-supervised video representation learning. Evolving loss functions have also been proposed to learn better video representations [157]. Transformers [204] have also made an impact on modeling video sequences, thus enabling cascading improvements in self-supervised video learning. The Timesformer architecture [20] has increased efficiency in processing large scale video data, and parameter efficient transformer models have also been proposed for video representation learning [109].

We draw inspiration from visual self-supervised learning for the learning of audio features. In this work, we apply concepts from visual self-supervision to develop two audio-only self-supervised methods (see Chapter 3) and a cross-modal self-supervised method based on visual reconstruction (see Chapters 4 & 5). Kolesnikov et al. [100] provide a holistic overview of visual self-supervised

learning for image representations for large scale data. The discussed pretext tasks include solving jigsaw puzzles (similar to [138, 99]), predicting image rotation (similar to [70]), and multi-task visual self-supervised learning [54]. The underlying principle of all of these ‘predictive’ self-supervised learning methods is very similar: “Predict a property or different part of the existing input data that requires the encoder network to learn a useful semantic property of the data”.



Figure 2.1: An illustration of the idea behind RotNet by Gidaris et. al. [70], where the goal is to learn unsupervised image features by image rotation prediction. The supervisory task is to select the image with the correct rotation out of 4 possibilities

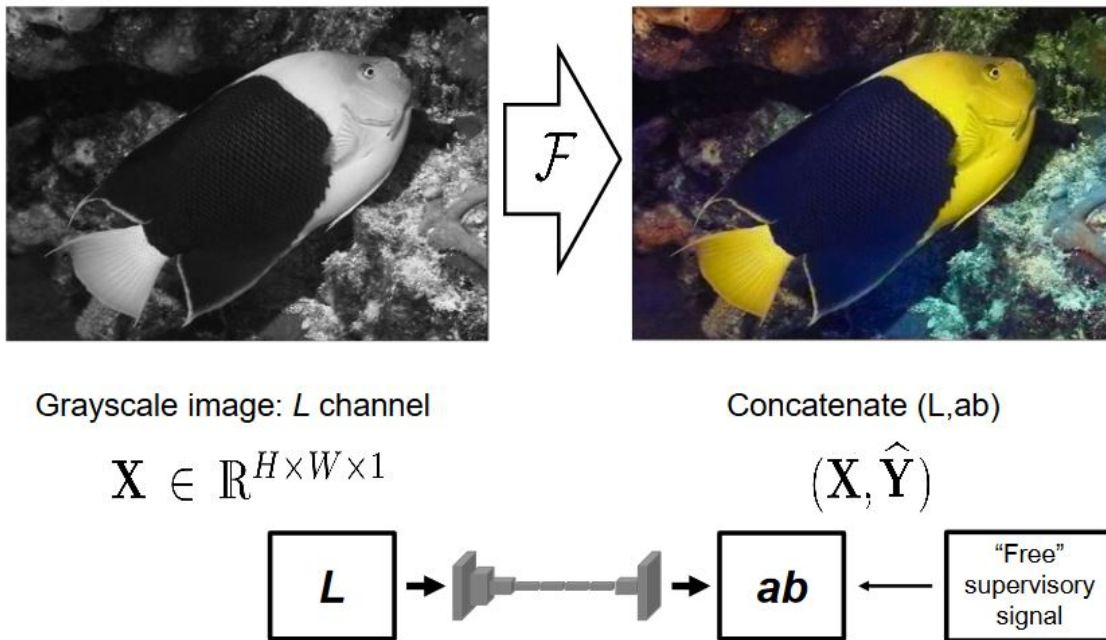


Figure 2.2: An illustration of the concept of colorization as a free supervisory signal from [205]. The prediction of the color channel values from the grayscale channel helps learn useful representations.

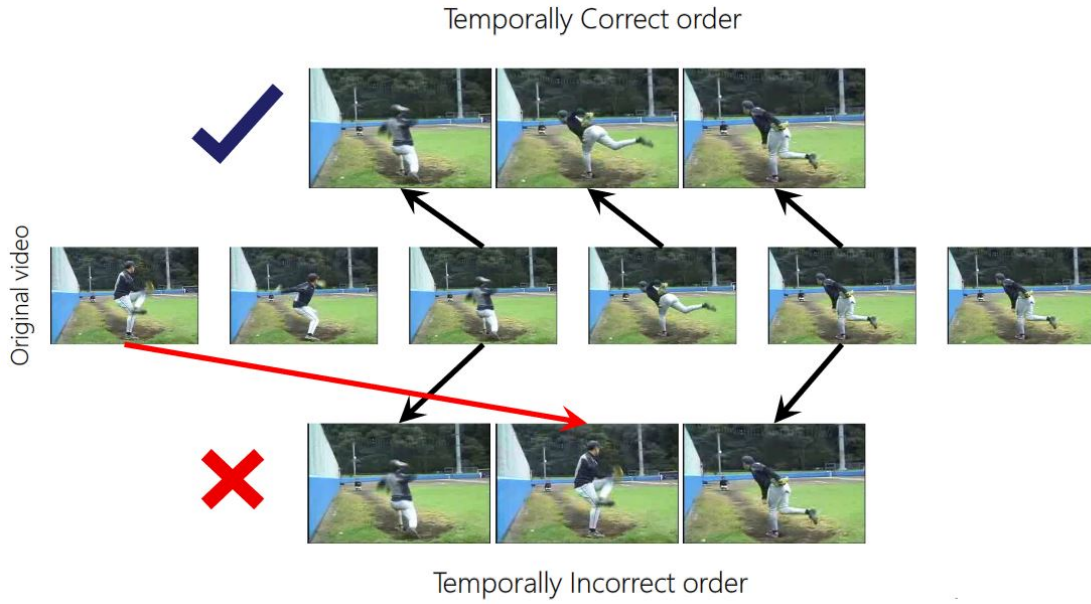


Figure 2.3: An illustration of Shuffle and Learn from Misra et al. [130], where the supervisory signal is temporal frame order verification. The input frames are presented in tuples of 3, with the positive examples being in correct order and the negative examples being in jumbled order. The task is to predict whether the presented tuple is in order or not.

2.3.2 Self-Supervised Learning for Text Data

Perhaps the most seminal and widespread recent applications of self-supervised learning have come in natural language processing. Extremely popular recent works like ELMo [149], GPT2 [162], and GPT3 [23] are based on predicting the next token of text based on the history or context. Models like BERT [53] and Roberta [114] were the first unimodal demonstrations of the currently very popular line of work of Masked Autoencoders (MAEs). Masked autoencoder based models have now gone on to find applications in other modalities (e.g. visual representation learning) and in multimodal settings (vision + text or audio + visual data). There was also prior work in learning vector representations from text such as word2vec [127] and fastText [21] which formed the cornerstones of modern unsupervised representation learning.

2.3.3 Self-Supervised Learning for Audio and Speech Data

Traditional audio representations were handcrafted; and often frequency domain; features like MFCC's (mel frequency cepstrum coefficients) and log mel spectrograms. There has been a wave of recent work on self supervised audio-only rep-

resentation learning which tries to automatically learn the best general purpose audio features for a wide number of tasks. We first examine contrastive learning based methods which model invariances between related parts of the audio signal to learn features that are similar to each other and distinct from other unrelated parts of the waveform. CPC (Contrast Predictive Coding) [140] and APC (Autoregressive Predictive Coding) [45] are similar approaches that model the next token of a speech segment given the history. Another method called LIM (Local Info Max) [165] is based on maximizing the MI (mutual information) among randomly chosen windows in an unsupervised way to learn speaker embeddings. Wav2vec [175] is an unsupervised pre-training method that builds on CPC in the context of speech recognition. Wav2vec [175] is a method trained by predicting speech units for masked parts of speech audio. Wav2Letter++ [159] is a similar idea for character based speech recognition. WaveNet [40] is a generative model for raw audio waveforms that can be used for generic audio representations. There has also been a new version of CPC proposed for audio for multiple languages [169]. Kharitonov et al. present an approach for contrastive learning in the time domain which is based on data augmentations [97]. This is complementary to a recent philosophical shift in some parts of the audio community to directly process raw waveform signals instead of frequency domain intermediate representations [164, 166, 137].

There is a different body of audio representation learning that is based on ‘predictive’ or ‘reconstructive’ based self-supervision. A very relevant recent work is PASE (Problem Agnostic Speech Encoder) [146] and its successor PASE+ [167], which aims to learn multi-task speech representations from raw audio by predicting a number of handcrafted features such as MFCCs, prosody and waveform. Phase prediction [161] has also been proposed as an audio-based pretext task.

The third distinct body of self-supervised learning methods for audio that we examine are Masked Autoencoders (MAEs). Two very popular recent examples are Wav2vec2 [15, 86] and HuBERT [85], which either use masked prediction as the pretext task or use pseudolabel based cluster prediction to iteratively refine the audio representations. Vector quantization has also been shown to be a useful step in learning better contextualized acoustic representations [112].

There are also some other interesting general directions in audio self-supervised learning. SeCoSt [107, 106] is a teacher-student self-supervised approach very similar to the ones in the visual domain that iteratively use the predictions of one epoch as the labels for the next one. Weakly labeled data can also help jointly learn both visual and audio representations [61]. Using two separate slow and

fast streams for audio recognitions (akin to the popular approach in video understanding) is also effective [96]. Self supervised audio features have also been proposed for mobile devices [195]. Xu et al. [218] show that self-training and pre-training are complementary for speech recognition. Baskar et al. [19] show how self-supervision can be used even on unpaired speech and text. Virtually all of these methods highlight the usefulness of using self-supervised learning to learn representations from unlabeled audio data (as opposed to existing supervised methods like Deepspeech2 [9]) to subsequently achieve better performance on downstream tasks of interest like speech recognition and acoustic classification [191].

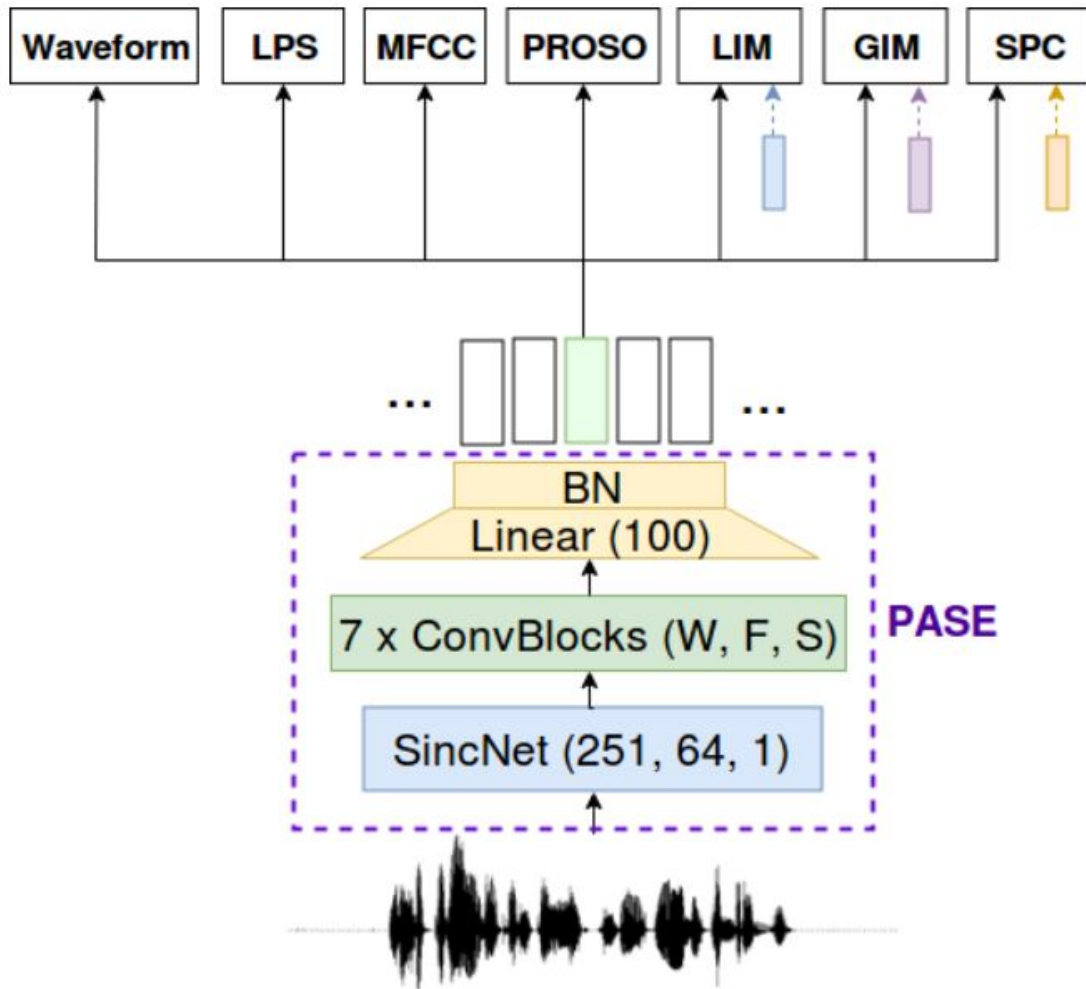


Figure 2.4: An illustration of PASE by Pascual et al.[146] taken from the Github repository. The idea is to predict a set of handcrafted features and properties from a raw audio waveform in order to make the encoder learn generalized and robust audio representations.

2.4 Multimodal Self-Supervised Learning

Thus far, we examined unimodal self-supervised learning and saw a wide variety of pretext tasks that are applicable to different kinds of data. We now present a review of multimodal self-supervised learning, which is able to utilize cross-modal correlations and complementary information across different modality pairs to learn even better representations. However, multimodal representation learning is not without its own set of quirks and challenges. Wang et al. [209] present a study about what makes training multimodal models difficult. There have also been strategies like curriculum learning [75] deployed to multimodal self-supervision which involve training with easier examples first and gradually increasing the difficulty over time.

The author would like to note a key theme here in multimodal self-supervised learning. There is great value in tasks like cross-modal generation, i.e. predicting one modality from another. We leverage this via speech-driven facial reconstruction later in the thesis. There is also a theme of cross-modal contrastive learning in which representations from multiple modality are brought closer in the latent space. However even if this produces stronger results, this is deeply unsatisfying because multimodal data is fundamentally complementary and heterogenous, and contrastive learning enforces homogeneity in the representation space ¹

We now review recent work in audiovisual self-supervised learning (which is the focus of this thesis) and also general cross-modal self-supervised learning.

2.4.1 Audiovisual Self-Supervised Learning

A few ways to utilize the relationship between modalities are by predicting audiovisual correspondence [10, 11], audiovisual synchronization [41, 141, 103], cyclic transitions [155], the relationship between ambient sound and vision [143], co-segmentation [171], and cross-modal prediction based fusion [150]. XDC [8] extends the idea of clustering as a pretext tasks across modalities, with the cluster predictions for video coming from audio and vice versa. Hu et al. propose a similar idea [87]. Piergiovanni et al. [157] propose a method that shares representations across modalities via distillation and finds better loss functions using evolutionary algorithms. Contrastive Multiview Coding [197] is a method that projects representations of views from different modalities closer in the latent space for positive examples and further for negative examples. In this process, the encoders

¹The author would like to acknowledge Andrew Owens for sharing this line of thought.

for each view (modality) learn useful representations. Patrick et al. [148] extend the contrastive learning concept from MoCo [81] to a multi-modal setting. There are also other works which use contrastive multimodal self-supervised learning to learn better audio and visual representations [6, 121, 12, 120]. Morgado et al. [134] combine audio-visual instance discrimination with cross-modal agreement for self-supervised learning. Zhu et al. [228] present a detailed survey of deep audio-visual learning including cross-modal self-supervised learning.

There is also an important line of work that focuses on learning representations for one modality with supervision from another. Owens et al. [143, 142] show that ambient sound provides supervision for visual learning. Aytar et al. in the Soundnet work show how it is possible to learn representations of sound that are guided by supervision from still images. The “2.5D Visual Sound” work by Gao et al. [65] shows how the visual modality can provide useful cues to predict a binaural output from a mono input audio. Morgado et al. [133] show how spatial audio can be generated using supervision from 360 video.

We now discuss work in self-supervised learning for audiovisual speech data which is a particular focus of this thesis. Ephrat et al. show an approach for speech separation that also uses visual information [58]. Joint audiovisual embeddings have also found creative uses. Chung et al. [44] show how discriminative audio and visual embeddings can be learned by cross-modal self-supervision. Nagrani et al. [136] show how cross-modal embeddings of both audio and video can be used as biometrics for authentication. It is also possible to learn the face behind a voice [139] and to generate the talking face of a person given their audio [42]. The latter is an idea that we use in great detail in this work to learn better audio representations.

Speech recognition (particularly audiovisual speech recognition) has also seen benefits from multimodal self-supervision [30, 145]. Important recent works for self-supervised audiovisual speech recognition (and visual-only speech recognition) include Lira [117], RAVeN [77], AV-HuBerT [178], and AV-Data2Vec [110]. These are all similar works that represent audiovisual speech using techniques like clustering, masked autoencoding, memory banks, and cross-modal prediction and they form the current state-of-the-art for self-supervised learning for audiovisual speech recognition.

Audiovisual sound source localization

We now review recent work in the related topic of audiovisual sound source localization. The fundamental problem being solved here is to try and determine the sound producing spatial (and spatiotemporal in video) regions in the visual modality for a given sound in the audio modality by leveraging the correct cross-modal associations. There is some historic work ranging from 1999 with “*Audio Vision*” by Hershey et al. [84] which proposed using audiovisual synchrony to locate sounds; to Fisher et al. [64] showing how audiovisual fusion and segregation can be jointly learned; and the 2005 work “*Pixels that sound*” by Kidron et al. [98]. In a slightly more modern context, Senocak et al. [176, 177] demonstrate this on paired image and audio, with supervision (even if weak) being shown as a critical component of the system to avoid learning spurious correlations. Tian et al. [198] built on the weak supervision idea for source localization. Qian et al [160] showed how multiple sound sources can be localized from a coarse to fine granularity. The “*Sound of Pixels*” and the “*Sound of Motions*” work Zhao et al [226, 227] showed how visual and motion information can be associated with the audio modality for sound localization. Izadinia et al. [92] show how segmentation of moving objects can be used for multimodal sound source localization. Tzinis et al. [201, 202] show very strong results on universal on-screen sound separation using audiovisual attention architectures.

There is also an emerging theme in this literature of trying to eliminate false positives (spurious correlations) and false negatives (missing out actual sound producing regions) by better modeling cross-modal correspondences and correctly accounting for cross-modal heterogeneity. Liu et al [113] propose a method based on cross-modal interference erasing, whereas Gao et al. [67] use cross-modal consistency. Gao et al. also study audiovisual co-separation [66].

Most works discussed so far operate using at least some form of supervision. We now discuss methods that use no labels whatsoever to try and achieve the same objective of localizing sound producing regions in the video. Gao et al. learn how to separate object sounds by learning cues from unlabeled videos [68]. In perhaps the most relevant work to this thesis from this area, Afouras et al. [5] in their *AVObjects* work use artificially misaligned negative audiovisual pairs to learn an audiovisual sound source localization map for general talking heads. There are also other works that do self-supervised audiovisual sound source localization based on exploiting cross-modal correspondences [88, 124, 39]. Chen et al. [34] perform sound object localization followed by trying to learn finer grained maps.

Mo et al. build on this idea in their EZVSL [132] and SLAVC [131] works, while proposing better evaluation metrics and achieving better performance for self-supervised audiovisual sound source localization.

There are also interesting new datasets that further facilitate research into this area. The VGG-SS dataset [33, 34] provides a larger scale audiovisual sound source localization benchmark whereas other datasets like the Soundspaces dataset [31] and the *AudioGoal* objective in Embodied AI literature extend the problem from 2D pixels in images and videos to simulated and real world navigation by embodied agents which is a very interesting area.

2.4.2 General Cross-Modal Self-Supervision

A very popular research area that we mention here for completeness but which is out of scope for this thesis is learning multimodal representations of vision + text data, which is very pervasive on the internet. Chen et al. [32] provide a survey of approaches for self-supervised vision-language pretraining. The underlying theory however, remains the same. Each modality (computational sequence) needs to be converted into a representation (feature vector of a certain dimension that represents the sequence) and there are multiple ways of self-supervised learning; invariance-based and generative; that offer ways to do this effectively and achieve good downstream performance.

There has been some recent work on developing general models that are able to represent any input modality using the same architecture. Of particular note are methods such as Perceiver IO [93], Unified IO [116], and Data2vec [16]. All of these methods use a transformer based encoder to convert input arrays of arbitrary dimensionalities to output representations of desired dimensionalities. These backbones can be trained in a self-supervised way by masked autoencoding and other modality specific self-supervision. These types of methods may pave way for a more general theory of simultaneously representing all kinds of input modalities in a shared and universal space, which is something that the author is deeply interested in pursuing in future work.

2.4.3 When does Multimodal Self-Supervision not work?

Multimodal Self-Supervised Learning definitely has a lot of promise when there are patterns across modalities that can be leveraged for representation learning (as is the case with audiovisual speech in this thesis which uses the correlation between

lip movements and speech content). However, we must be careful to not always assume that cross-modal patterns exist or are useful. Multimodal information is often a mixture of *redundant* and *complementary*. If the *complementary* part of the information is too high (or is *conflicting*), then approaches like contrastive learning or cross-modal reconstruction will not be able to use a proper supervisory signal to learn representations from. Multimodal correlations can also often be *spurious*. For example, in driving videos you will usually always see the sky in addition to the body of the car. Incorrectly modeling cross-modal correlations (e.g. for audio and video) could lead to the model learning to associate the car’s engine sound with the visual content of the sky (which is clearly a spurious correlation).

2.5 Applications of Self-Supervised Learning for Audio and Speech

We now briefly discuss application areas in audio and speech that can especially benefit from self-supervised learning.

2.5.1 Speech Recognition and Enhancement

Speech recognition is perhaps the most important human-centric audio recognition problem, and multiple self-supervised learning approaches have demonstrated significant results on it. Of particular relevance here is speech recognition for languages and settings that have limited labeled data. Self-supervised learning has been shown to be extremely effective in situations like this (which is also something we explore in later chapters of this thesis). Multimodal self-supervision in audiovisual data can also be used to drive better audio representations for audio-only speech recognition. Speech enhancement is also an important problem which self-supervised learning (and particularly visually guided self-supervised learning) can help. By learning the implicit distribution of clean and noisy data (e.g. by artificially creating noisy data), we can improve downstream recognition performance by using representations that are more robust. The visual modality can help learn features that are more robust in cases where auditory noise is present.

2.5.2 Affect/Emotion Recognition

Emotion recognition; both from discrete categorical labels and continuous valence/arousal labels; is a problem that is usually accompanied by small and spe-

cific labeled datasets on which it is difficult or intractable to train supervised models from scratch. It is also usually expensive to find skilled annotators to provide gold standard annotations for this kind of data (especially for things like pain estimation). In these situations, self-supervised learning has tremendous potential to make improvements relative to existing methods. Cross-modal self-supervision may also help learn complementary cues across modalities; e.g. video has facial expression information which map to speech emotion in the audio, and can be used to drive better downstream audio emotion recognition. We explore downstream emotion recognition in great detail in this thesis.

2.5.3 Audio Event Detection

Understanding sounds in the natural world is an important problem to determine the context of an artificially intelligent agent. The audio modality has a lot of useful information of many different kinds (e.g. emergency vehicle sirens, music understanding, recognizing the specific environment and its acoustics). All of these types of data are difficult to label at a fine grained level at scale, and self-supervised learning is appropriate to learn better general purpose representations of sound. Datasets like Audioset [69], Urban Sounds [173], and ESC50 [156] are examples of tasks that will benefit tremendously from audio and audiovisual self-supervised learning.

2.5.4 Other Audio Applications

There may be a lot of other interesting audio applications that benefit from self-supervised learning. A few specific ones that are notable for audiovisual self-supervision are egocentric audiovisual understanding for datasets like Ego4d [72], Epic kitchens [48], and Epic sounds [89]. Another very interesting line of research is audio understanding in Embodied AI [51], where large-scale acoustic understanding in simulations can be transferred to real-world systems.

Chapter 3

Audio-only Self-Supervised Learning of Speech Representations

The contents of this chapter are partially based on the publication “Does Visual Self-Supervision Improve Learning of Speech Representations for Emotion Recognition?” in IEEE Transactions on Affective Computing, 2021 [188]. In this chapter, we introduce two new methods for learning audio representations: (i) Arrow of Time, (ii) Odd One Out. Both of them are based on predicting the temporal order of a sequence of audio.

The main focus of this thesis is to learn better representations of audio. The audio modality contains a lot of important information about human-centric communication, most notably speech content. Given the scale of audio data that modern machine learning systems need to accurately process, self-supervised learning of audio representations is a crucial area in which making advances would have a cascading and significant positive effect on a lot of applications. A distinguishing factor for audio is that the modality is very very efficient at encoding salient information about speech content from an information theoretic perspective. In one seconds worth of audiovisual speech data, with a 25 fps 224x224x3 video modality and 16000kHz audio modality, the video tensor is of a shape 25x224x224x3 which is 3,763,200 total values, as opposed to the audio signal which only has 16,000 values. This is a factor of over 200x the amount of data needed to represent the signal as a tensor. Despite being lesser total data overall, audio usually contains a clearer signal of speech content, whereas video contains some redundant and other complementary information (e.g. facial expressions). Thus, it would make sense to learn self-supervised representations of the audio modality for speech related problems.

However, the visual modality has seen far more attention and research contribution in the area of self-supervised learning than audio. In this thesis, we try and narrow this gap by developing methods that can yield better audio representations that are generally useful for a variety of speech and general sound tasks.

The raw audio signal is essentially a waveform that is a 1-D tensor with a specified sampling rate. In this thesis, we only work with single channel audio. However, any of the existing or proposed theory can be extended to multiple audio channels by treating them as such. Most of the useful information in the audio signal for tasks of interest is present in its frequency over time. Traditional representations such as MFCC's (cepstral-coefficients of mel frequencies) and log mel spectrograms have been widely used and optimized for speech related tasks over the years and offer an extremely strong baseline. They are both computed using the Mel scale which is a non-linear transformation of the frequency scale to make it maximally discriminative to human perception of frequencies. They have also been treated equivalent to images of [frequency, time] dimensions and processed using architectures optimized for visual processing in recent representation learning literature.

Despite the existence of strong supervised learning systems for speech recognition and other speech related tasks, self-supervised learning for audio has been relatively underexplored. In this chapter, we propose two self-supervised methods to learn better audio speech representations. These methods operate on log mel inputs and are inspired by temporal order verification based pretext tasks that were first demonstrated in computer vision. We demonstrate the effectiveness of these methods on speech recognition and speech emotion recognition.

3.1 Background

While we have introduced related work in audio self-supervised learning in Chapter 2, we now briefly (re)introduce the audio-only self-supervised learning methods we use as baselines for this work in greater detail.

There is also other recent work that is based on contrastive learning such as CPC [140]. The method operates on raw audio and is trained by contrasting near future audio windows from other audio windows that are either far away in temporal distance or come from different audio samples altogether. CPC tries to model a density ratio to maximize mutual information (MI) between the target signal (random raw audio window) and the context (current raw audio window).

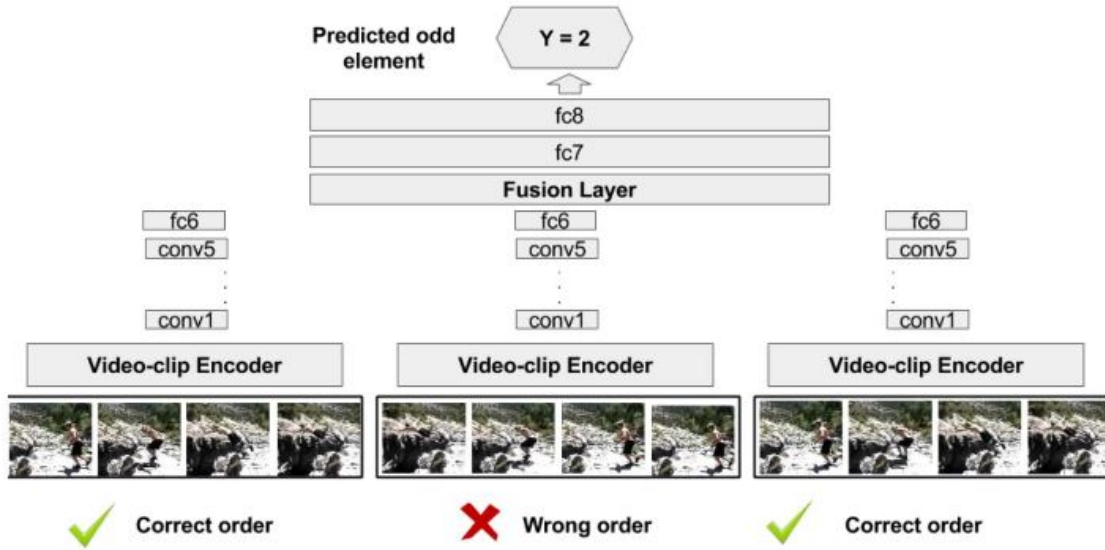


Figure 3.1: An illustration of Odd One Out networks from Fernando et al. [63], where the supervisory signal is temporal frame order verification. The input video is jumbled in one of the input sequences and the task is to predict which one out of the inputs is the jumbled element. We implement an audio adaptation of this method.

By maximizing the MI, the method can extract the underlying latent variables that the two different parts of the signal have in common. This essentially means that the representations of related (positive) examples lie closer in the latent space, and those of negative examples lie further away. CPC has also been shown to generalize well across languages [169]. wav2vec [175] refines the idea of CPC for speech signals.

Autoregressive Predictive Coding (APC) [45] is similar to CPC, however the key difference is that APC directly tries to predict the immediate future part of the signal based on the history whereas CPC tries to maximize mutual information between the target (future) and the context (present). The input features for APC are 80 dimensional log mel spectrograms with a window size of 25 ms and a step size of 10 ms. The model tries to predict the log mel spectrograms for the future windows given the history.

PASE [146] is a raw audio encoder trained in a self supervised way to predict various different handcrafted features such as MFCC, prosody, waveform etc. While predicting these multiple tasks, the encoder learns a very robust and multi-task representation for raw audio that these tasks exemplify (e.g. prosody for emotion). This is a form of generative / reconstructive self-supervision in which the input signal (raw audio) is used to generate a variety of outputs (speech attributes) that are then compared against the ground truth speech attributes to

train the model.

All of these methods offer strong baselines for speech representations and differ in their inspiration, execution, and operation (e.g. raw audio inputs vs log mel inputs).

3.2 Audio Self-Supervision for Speech Representation Learning

This section introduces the audio-only self-supervised methods that we propose for speech representation learning. The concept behind both is temporal order verification for audio as a pretext task. The methods are inspired by similar works in visual self-supervision called ‘Arrow of Time’ [215] ‘Odd One Out’ [63].

3.2.1 Arrow of Time for Audio

The temporal order of a sequence carries a lot of potentially useful information about its structure. For video, Wei et. al. [215] proposed the Arrow of Time as a self supervised method that predicts whether a given video sequence is being played forwards or backwards. This helps the encoder that is predicting this pretext task label to learn features that correspond to object semantics and other visually correlated physical characteristics like gravity, forces etc. which may be useful for generic visual feature learning. We adapt the Arrow of Time (henceforth abbreviated as AoT) method for speech signals. The problem reduces to predicting whether a given audio clip is being played forwards or backwards. While learning to predict the task, the encoder learns useful audio features that differentiate between certain phonemes and how they sound when played forward vs backward. In order to predict the direction of the Arrow of Time, the encoder must capture useful characteristics about the phonemes themselves. In our implementation, we simply flip the temporal order of half of the sequences of an input batch (make them play backwards), and train the audio encoder with the supervised task of predicting the binary class problem (forward or backward). We use the encoder architecture described in Section 3.2.4. The experimental setup and results are discussed in Section 3.3.

3.2.2 Odd One Out for Audio

While predicting the forwards or backwards direction of the sequence is definitely valuable, there may be more refined pretext tasks that are also based on predicting other aspects of the temporal order of a sequence. Odd One Out networks for video [63] are based on predicting which one out of multiple sets of ordered sequences of frames is in jumbled order (temporally incorrect order). This helps the encoder that is predicting this pretext task label to learn features that correspond to object semantics and other visually correlated physical characteristics. We adapt this idea to the audio modality in a straightforward way. For a given input batch of audio clips, we jumble 25% of the clips. The jumbling is performed by selecting at random two windows of a length of 15% of the total audio duration and swapping them. The encoder is then tasked with predicting which element in the input batch is the 'Odd One Out', and is optimized using cross entropy loss. While learning to predict the task, the encoder learns useful audio features that differentiate between certain phonemes and how they sound when played jumbled vs in-order. In order to predict whether the input speech sequence is jumbled, the encoder must capture useful characteristics about the phonemes themselves. Fig. 3.2 illustrates the training procedure for Odd One Out networks for audio representation learning. We use the encoder architecture described in Section 3.2.4. Our results are described in Section 3.3

3.2.3 Reconstructing Salient Speech Attributes

We also propose another method for audio-only self-supervised learning of speech representations in Chapter 6. We mention this here for completeness, however the details and results for that specific method is presented in better context in Chapter 6. The fundamental difference between the methods presented here and that which is presented later is that Odd and AoT operate on log mel inputs (which are already a good starting point as audio features), whereas the reconstruction based method operated directly on raw audio.

3.2.4 Audio Encoder Architecture

The audio encoder is a log mel spectrogram encoder (closely following [45]). The log mel spectrogram is computed with 80 frequency bins, a window width of 25ms and a stride of 10ms, which is a standard choice for processing speech signals. This (t, 80) dimensional input then goes through the encoder which is a 3 layer GRU

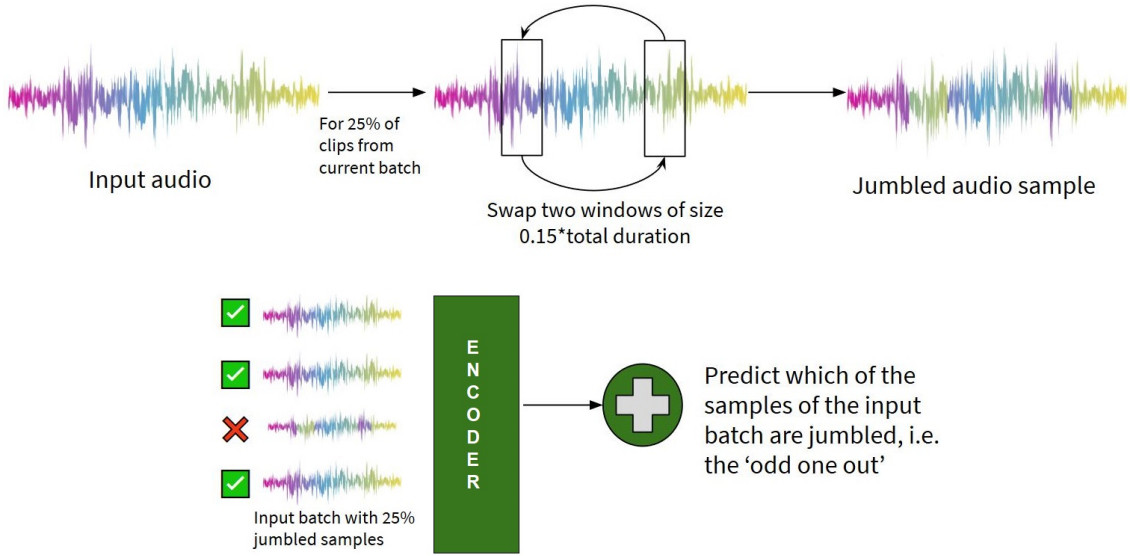


Figure 3.2: An illustration of our proposed variation of Odd One Out networks for audio. The supervisory signal is temporal audio frame order verification. The input audio sample for 25% of the input batch are jumbled. The jumbling is performed by swapping two random windows each of 15% of the total duration of the clip. The audio encoder is then trained to predict which ones out of the input batch are jumbled or the ‘odd one out’.

network with each layer having a hidden size of 512 followed by a fully connected layer which converts it into a feature with dimensionality $(t, 512)$. This specific architecture of the audio encoder with 3 GRU layers is the exact same as used in [45]. We chose this for simplicity and to enable direct comparison with this baseline which uses a similar audio input as us (80 dimensional log mel features).

We use the above described architecture as the audio encoder in the proposed models in this chapter. We then use the trained encoder to extract features from the evaluation datasets. We implement our method and all evaluation pipelines using PyTorch [147].

3.3 Experiments and Results

3.3.1 Datasets

This section introduces the various audio-only and audiovisual datasets that were used in the work either for pretraining or evaluating the baseline and proposed models. For all datasets, we divide the data into training, validation and test sets

with all samples from each speaker belonging to a particular set only. Table 5.1 summarizes the statistics for all the datasets used in this work.

The CREMA-D dataset [25] contains a diverse set of 91 actors who utter 12 sentences multiples times each with a different level of intensity for each of 6 basic emotional labels (anger, fear, disgust, neutral, happy, sad). We use Crema as a discrete emotion recognition evaluation dataset.

The Ravdess dataset [115] contains 1440 samples of 24 different actors who acted out two sentences with 8 different basic emotions (anger, calm, sad, neutral, happy, disgusted, surprised, fear) and two different intensity levels. We use Ravdess also as a discrete emotion recognition evaluation dataset.

The RECOLA dataset [168] contains dyadic conversations in French between a pair of participants working to solve a collaborative task over video conference. The annotated part of the dataset consists of 5-minute long clips (with audio from one speaker only) with continuous valence and arousal annotations from 6 annotators. Since the labels of the test set are not publicly available, we report the performance of our models on the development set.

The SEWA dataset [104] contains dyadic conversations over video conference between a pair of participants discussing about an advertisement that they have just watched. The audio clips are typically 3-minute long and have continuous valence and arousal annotations. We use the AVEC2017 challenge German language subset (with the gold standard valence and arousal labels) of the dataset to run our experiments.

The IEMOCAP dataset [24] contains dyadic conversations between 10 speakers for a total of 12 hours of audiovisual data. The discrete emotion labels comprise of 8 categories (anger, happiness, sadness, neutral, excitement, frustration, fear, surprise), however we only consider the first 4 categories for our experiments (anger, happiness, sadness, neutral). This is due to much higher inter annotator agreement for these categories, and this portion of the dataset has been similarly used in prior studies [122]. This partition also leaves us with around 6.5 hours of data instead of the original 12 hours. We use IEMOCAP as another discrete emotion recognition evaluation dataset.

The SPC (Speech Commands) dataset [212] contains 64,727 total utterances of 30 different words by 1,881 speakers. We use SPC as a speech recognition evaluation dataset.

The LRW dataset [41] is a large, in-the-wild dataset of 500 different isolated words primarily from BBC recordings. It is an audiovisual speech dataset and is thus appropriate for training our methods. We use a subset of LRW that has

Downstream Results		Discrete Emotion		
Pretraining Dataset		LRW	LRW	LRW
Evaluation Dataset		CREMA-D	Ravdess	IEMOCAP
Labels		6 emotions	8 emotions	4 emotions
Method	Pretext task	F1 Score ($\uparrow$)	F1 Score ($\uparrow$)	F1 Score ($\uparrow$)
MFCC	None	0.447 ± 0.012	0.390 ± 0.013	0.535 ± 0.006
Log Mel	None	0.463 ± 0.013	0.480 ± 0.010	0.572 ± 0.011
LLD [59]	None	0.409 ± 0.015	0.398 ± 0.013	0.497 ± 0.013
CPC [140]	Audio	0.422 ± 0.016	0.408 ± 0.011	0.517 ± 0.011
PASE [146]	Audio	0.488 ± 0.010	0.472 ± 0.010	0.557 ± 0.017
APC [45]	Audio	0.480 ± 0.012	0.497 ± 0.011	0.592 ± 0.019
AoT	Audio	0.535 ± 0.016	0.497 ± 0.008	0.602 ± 0.010
Odd	Audio	0.542 ± 0.016	0.501 ± 0.008	0.615 ± 0.010
Supervised	None	0.540 ± 0.013	0.528 ± 0.010	0.623 ± 0.010

Table 3.1: Results for all baseline and proposed methods for discrete emotion recognition (on CREMA, Ravdess and IEMOCAP). Unless specified otherwise, all methods are used as frozen audio feature extractors before training a classifier (BGRU) on the downstream task. All results are in the form of $\mu \pm \sigma$ for 10 runs.

only nearly frontal videos (with yaw, pitch and roll restricted to a maximum of 10 degrees), in order to have a cleaner supervisory signal from the visual modality. This filtering leaves us with a total of around 40 hours of usable data. We use LRW as the self-supervised pretraining dataset for all baseline and proposed methods.

3.3.2 Baselines

Since the aim of our work is to yield self-supervised audio features, we compare against other baselines focusing on the same goal. The three self-supervised methods we compare against are CPC [140], APC [45] and PASE [146].

We also compare our methods against 39 dimensional MFCCs (13 coefficients, 13 deltas, and 13 delta-deltas) which act as baseline features used for supervised learning for audio. We also compare with 80 dimensional log mel spectrograms. Additionally, we also use the OpenSmile [59] toolkit to extract LLD’s (low level descriptors) based on the IS13-ComParE feature set. These features are well established baselines for audio based affective computing tasks.

Downstream Results		Continuous Affect Recognition	
Pretraining Dataset Evaluation Dataset		LRW SEWA	
Labels		Valence	Arousal
Method	Pretext task	CCC ($\uparrow$)	CCC ($\uparrow$)
MFCC	None	0.349 ± 0.009	0.359 ± 0.010
Log Mel	None	0.307 ± 0.011	0.316 ± 0.007
LLD [59]	None	0.245 ± 0.011	0.236 ± 0.010
CPC [140]	Audio	0.287 ± 0.008	0.295 ± 0.009
PASE [146]	Audio	0.354 ± 0.008	0.343 ± 0.010
APC [45]	Audio	0.317 ± 0.009	0.321 ± 0.014
AoT	Audio	0.339 ± 0.006	0.340 ± 0.009
Odd	Audio	0.342 ± 0.008	0.339 ± 0.010
Supervised	None	0.297 ± 0.019	0.330 ± 0.009
Downstream Results (contd.)		Continuous Affect Recognition	
Pretraining Dataset Evaluation Dataset		LRW RECOLA	
Labels		Valence	Arousal
Method	Pretext task	CCC ($\uparrow$)	CCC ($\uparrow$)
MFCC	None	0.413 ± 0.008	0.714 ± 0.011
Log Mel	None	0.396 ± 0.012	0.696 ± 0.011
LLD [59]	None	0.271 ± 0.015	0.624 ± 0.010
CPC [140]	Audio	0.320 ± 0.011	0.623 ± 0.012
PASE [146]	Audio	0.390 ± 0.008	0.719 ± 0.007
APC [45]	Audio	0.383 ± 0.010	0.703 ± 0.008
AoT	Audio	0.431 ± 0.011	0.719 ± 0.008
Odd	Audio	0.436 ± 0.012	0.720 ± 0.008
Supervised	None	0.402 ± 0.010	0.701 ± 0.008

Table 3.2: Results for all baseline and proposed methods for continuous affect recognition (on SEWA and RECOLA). Unless specified otherwise, all methods are used as frozen audio feature extractors before training a classifier (BGRU) on the downstream task. All results are in the form of $\mu \pm \sigma$ for 10 runs.

Results		Speech Recognition
Pretraining Dataset		LRW
Evaluation Dataset		SPC
Labels		30 words
Method	Pretext task	F1 Score ($\uparrow$)
MFCC	None	0.912 ± 0.002
Log Mel	None	0.927 ± 0.003
LLD [59]	None	0.643 ± 0.014
CPC [140]	Audio	0.814 ± 0.019
PASE [146]	Audio	0.941 ± 0.002
APC [45]	Audio	0.933 ± 0.003
AoT	Audio	0.917 ± 0.002
Odd	Audio	0.919 ± 0.002
Supervised	None	0.921 ± 0.005

Table 3.3: Results for all baseline and proposed methods for speech recognition (on SPC). Unless specified otherwise, all methods are used as frozen audio feature extractors before training a classifier (BGRU) on the downstream task. All results are in the form of $\mu \pm \sigma$ for 10 runs.

3.3.3 Experimental Setup

This section presents the details of all experiments that we perform to rigorously validate our proposed methods. We present all results for discrete emotion recognition (Table 3.1), continuous affect recognition (Table 3.2), and speech recognition (Table 3.3) from the extracted features.

We evaluate all extracted features on: (i) Discrete Emotion Recognition, (ii) Continuous Affect Recognition, and (iii) Automatic Speech Recognition (ASR).

For the **emotion recognition** task, we first perform self-supervised pretraining on LRW as described, and then use the pretrained models as feature extractors on the CREMA, Ravdess and IEMOCAP datasets. Once we have these features, we then train an BGRU (bidirectional gated recurrent unit) model for the emotion classification task. We used a BGRU for simplicity, however this can be replaced by any model that can classify variable length sequences into discrete categories (such as LSTMs, TCNs, Transformers, LiGRUs [166]). We use a 2 layer BGRU with 256 units in each layer. The initial learning rate is 0.0001 and is decayed by a factor of 0.1 every 40 epochs. We train for 100 epochs and use the checkpoint from the epoch which gives the best validation set metric for evaluation on the test set. We pass the last hidden state of the BGRU to a linear layer with size equal to the number of target classes (6 for CREMA, 8 for Ravdess, 4 for IEMOCAP) followed by a Softmax layer with a cross entropy loss for emotion classification. This process (self-supervised feature extraction + BGRU training) is followed for all compared methods (see bottom of Fig. 5.1).

For the **speech recognition** task, we use the SPC dataset to evaluate our method. For the SPC dataset which is a spoken word classification task with 30 different possible labels, we use the exact same protocol as described for emotion recognition (self-supervised feature extraction + BGRU training). We use the same parameters and learning schedule for the BGRU.

For the **continuous affect recognition** task, we use SEWA and RECOLA as the evaluation datasets. We predict continuous values of valence and arousal for each audio segment. We train separate individual models for valence and arousal prediction. Using the same encoder architecture and self-supervised pretraining followed by downstream evaluation process described in the bottom of Fig. 5.1, we use the output of the BGRU at each timestep as the predicted valence or arousal value. This gives us a predicted sequence y . We train the network to maximize the Concordance Correlation Coefficient (CCC) between the predicted sequence y and the ground truth valence or arousal sequence $\hat{y}$. The formula for CCC is:

$$CCC = \frac{2\sigma_{y\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2} \quad (3.1)$$

where μ_y and $\mu_{\hat{y}}$ are the means of the ground truth and predicted sequences, σ_y^2 and $\sigma_{\hat{y}}^2$ are the variance of the ground truth and predicted sequences, and $2\sigma_{y\hat{y}}$ is the covariance between the ground truth and predicted sequence. The value of CCC lies between -1 and 1, and is 0 for uncorrelated sequences. The loss function that we minimize to train our network (which maximizes CCC) is $1 - ((CCC+1)/2)$, which lies between 0 and 1. For SEWA, we report the CCC on the test set as the evaluation metric. For RECOLA, we report the CCC on the validation set to because test set labels are not publicly available. Note that CCC is the standard metric for the SEWA dataset (used in the AVEC 2017 challenge) and the RECOLA dataset (used in the AVEC 2016 challenge).

For all results in Tables 3.1, 3.2, and 3.3, we validate the models using the validation set of the downstream task and use the model from the epoch with the best validation performance to evaluate on the test set. We report the results in the form of $\mu \pm \sigma$ (mean $\pm$ std) over 10 runs for each experiment.

3.3.4 Results

Our methods for audio-only self-supervision (Odd One Out and Arrow of Time) are indicated as **Odd** and **AoT** throughout the discussion of the results.

For discrete emotion recognition, Odd is the best performing audio-only self-supervised method on CREMA (F1 score of 0.542), Ravdess (F1 score of 0.501) and IEMOCAP (F1 score of 0.615). For continuous affect recognition on SEWA, Odd achieves a valence CCC of 0.342, which is outperformed by MFCCs with CCC of 0.349 ($p=0.13$). For arousal, Odd achieves CCC of 0.339 which is outperformed by both MFCCs and PASE with CCCs of 0.359 ($p<0.01$) and 0.343 ($p=0.362$) respectively. On RECOLA, Odd is the best method for arousal (CCC=0.720) and valence (CCC=0.436). For speech recognition on SPC, both PASE (F1 score of 0.941) and APC (F1 score of 0.933) significantly ($p<0.01$) outperform Odd (F1 score of 0.919).

We notice that while Odd is the best method seen so far for discrete emotion and offers competitive performance with respect to other methods on the other tasks, there is a performance gap for speech recognition and continuous affect recognition with MFCCs and PASE. We aim to bridge this performance gap and yield better features using visually guided self-supervised learning in Chapter 5

(as opposed to the audio-only pretext task in this chapter).

When comparing between the two proposed methods, Odd and AoT seem to be very close in performance on emotion recognition, but Odd seems to slightly outperform AoT on average on most tasks (likely due to being a more refined pretext task). Both methods outperform other baselines for audio-only self-supervision for emotion recognition in most settings. However, APC and PASE prove to be more effective for speech recognition.

3.4 Discussion

The proposed methods for audio-only self-supervision are effective for emotion recognition where they outperform the baselines, but the baselines are marginally better for speech recognition. This points to the pretext task being useful for learning emotion representations (the frequency domain is very useful for emotion and the methods are ‘enhancing’ the already good log mel inputs), but some additional form of self-supervision being needed to potentially learn even better representations that are especially helpful for speech recognition.

While predicting the temporal order of the sequence is an interesting pretext task, in the next chapters we leverage the visual modality as well. Of particular interest is the ability to leverage the auxiliary information present in lip movements in the video modality to drive the learning of better audio representations (which is something that has not been explored in the past). Having access to the visual modality will also provide a source of complementary information which audio alone simply cannot do. In subsequent chapters, we explore visual and joint audiovisual self-supervision with the aim to improve the performance achieved by the currently proposed audio-only self-supervision method.

Chapter 4

Visually Guided Self-Supervised Learning of Speech Representations

The contents of this chapter are based on the publication “Visually Guided Self Supervised Learning of Speech Representations” in the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2020 [187]. In this chapter, we introduce the core idea of this thesis: using visual self-supervision from lip movements to guide the learning of better audio representations. We use speech-driven facial reconstruction as a self-supervised learning pretext task, and show that it is effective for audio feature learning.

4.1 Background

We discussed a wide variety of audio-only and audiovisual methods for self-supervised learning of audio features in Chapter 2. All of these works have shown that it is possible to learn robust multi-task representations from a large amount of unlabeled data that is inexpensive to obtain. In the previous chapter we also saw how a proposed audio-only self-supervised method is able to achieve competitive results (and in some cases outperform baselines) on speech recognition and speech emotion recognition. To further improve the performance, we also introduce the visual modality into the process for self-supervised learning. We propose a novel self-supervised method based on cross-modal reconstruction to learn audio features. Our method is based on speech-driven facial reconstruction. Given an input frame and the corresponding audio, we generate/reconstruct the ground truth talking face video sequence. Speech based facial reconstruction has a long history in computer vision, and has been demonstrated by works like Video Rewrite (1997) [22],

Mary 101 (2002) [60] and more recently Synthesizing Obama (2017) [194]. Our contribution is to utilize **speech-driven facial animation as a pretext task for self-supervised learning of audio features**, which has not been explored before. We exploit the correlation between facial expressions and emotion, and that between lip movements and speech content. This leads to audio features that contain additional information about emotion and speech due to using the visual modality as a supervisory signal.

The current state of the art for both affect and speech recognition is largely dominated by fully supervised models which require large amounts of labeled training data. However extensive annotation is difficult and expensive, especially for problems like emotion recognition. Self-supervised learning may offer a practical solution to this issue by virtue of being able to learn from large amounts of unlabeled data.

Most existing self-supervised learning approaches are either unimodal (for audio and video). The few cross-modal approaches typically have some interaction between the modalities in the latent space by pretext tasks like clustering, however they don't produce an interpretable and intuitive interaction between the two modalities. Our work proposes audio features that are explicitly guided by the generation of lip movements and facial expressions by reconstruction. Our proposed method is novel in its methodology and occupies an interesting niche in audio representation learning. We capture explainable visual information in the audio features, however we can evaluate our model on audio-only datasets without requiring the visual modality. Our work highlights the ability to use visual self-supervision using any audiovisual speech dataset to improve the performance of any audio-only method on a variety of problems and under various levels of noise.

4.2 Speech-Driven Facial Video Reconstruction for Self-Supervised Learning

The proposed architecture is shown in Fig. 4.1 and is based on our prior work on speech-driven facial animation [206]. The model is a temporal encoder-decoder which takes a still image (frame from a 25 fps video) and an audio signal as input. The audio (16 kHz waveform) is divided into overlapping windows of 200 ms, with each window centered around a video frame. The model itself can be divided into three subnetworks as shown in Fig. 4.2, namely the content encoder (6

layer 1D CNN audio encoder + GRU), the identity encoder (6 layer 2D CNN) and the frame decoder (U-Net [170] architecture with skip connections from identity encoder, layer sizes and parameters are same as U-Net).

The audio frame encoder (see bottom of Fig. 4.2) converts an 3200x1 audio window into a 256-dimensional feature vector z_{aud} as shown. Similarly, the identity encoder, which is made of 6 (Conv2D - BatchNorm - ReLU) blocks, reduces a 96x128 input image to a 128x1 feature vector z_{id} . We also use a Noise Generator capable of producing noise that is temporally coherent. A 10 dimensional vector is sampled from a Gaussian distribution with mean 0 and variance of 0.6 and passed through a single-layer GRU to produce the noise sequence. This latent representation z_n accounts for randomness in the face synthesis process. The latent representation is the concatenation of z_{aud} , z_{id} and z_n . This embedding then goes through the frame decoder, which is a CNN that uses strided transposed convolutions to produce the video frames. The skip connections to the identity encoder help in preserving subject identity.

An L1 reconstruction loss between a random frame from the generated video and the corresponding frame from the real video is used to train the network. We use the Adam optimizer with a learning rate of 0.06 that is decayed by a factor of 0.98 every 10 epochs. The network learns to reconstruct the face. In doing so, the audio encoder is driven to produce useful speech features that contain information about mouth and facial movements. These representations can then be used for downstream tasks like ASR and emotion recognition.

4.2.1 Why Facial Reconstruction is Relevant for Emotion and ASR

Facial reconstruction provides a very explicit cross-modal correspondence between audio and video in audiovisual speech data. Of particular interest are the correlation between **lip movements and speech content**, and the correlation between **facial expressions and speech emotion**. If we can successfully reconstruct these cross-modal correlations by the speech-driven facial reconstruction process (i.e. by reducing the reconstruction loss), then the audio encoder *must* contain useful representations of speech content and speech emotion. If it does not successfully represent these while being the latent variable that the reconstruction process is conditioned on, then the reconstruction cannot realistically contain useful information about these. Doing the reconstruction successfully enables us to learn better audio representations in general, and better representations for emotion

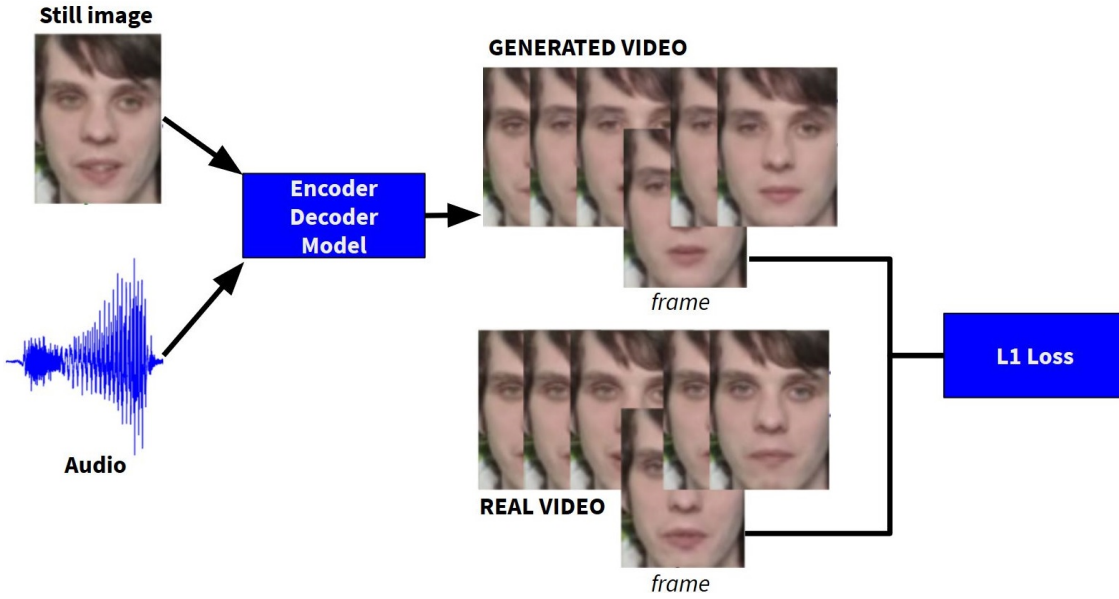


Figure 4.1: An overview of our proposed model. We generate a video from a still face image and the corresponding audio and optimize the reconstruction loss. In order to reconstruct the face correctly, the model needs to capture lip movements. This drives the learning of better audio features.

Dataset	Train	Val	Test
GRID	31639 / 26.4	6999 / 5.80	9976 / 8.31
TCD TIMIT	8218 / 9.10	686 / 0.80	977 / 1.20
LRW	112658 / 36.3	5870 / 1.90	5980 / 1.90
CREMA-D	11594 / 9.70	819 / 0.70	820 / 0.68
Ravdess	1509 / 1.76	415 / 0.48	519 / 0.60
SPC	51094 / 14.2	6798 / 1.88	6835 / 1.89

Table 4.1: The samples and hours (number, time in hours) of audiovisual speech data in the training, validation and test sets of each dataset.

and speech recognition in particular.

4.3 Experimental Setup

4.3.1 Datasets

This section introduces the various audiovisual datasets that were used in the work either for pretraining or evaluating the baseline and proposed models. For all datasets, we divide the data into training, validation and test sets with all samples from each speaker belonging to a particular set only.

The GRID dataset [47] contains audio-visual speech recordings of subjects with

a frontal view. It has 33 speakers, each of whom speak 1000 sentences containing six words. We use GRID as an ASR evaluation dataset. The TCD TIMIT [79] dataset contains 59 speakers uttering 100 phonetically rich sentences sourced from the original TIMIT ASR dataset. We use TCD TIMIT as a pretraining dataset. The LRW dataset [41] is a large, in-the-wild dataset of 500 different isolated words primarily from BBC recordings. We use a subset of LRW that has only nearly frontal videos (with yaw, pitch and roll restricted to a maximum of 10 degrees). We use LRW as a large sized pretraining dataset. The CREMA-D dataset [25] contains a diverse set of 91 actors who utter 12 sentences multiples times each with a different level of intensity for each of 6 basic emotional labels (anger, fear, disgust, neutral, happy, sad). We use CREMA-D for both pretraining and evaluation for emotion recognition, but not for ASR because it is phonetically very limited even though it is larger than TCD TIMIT. The Ravdess dataset [115] contains 1440 samples of 24 different actors who acted out two sentences with 8 different basic emotions (anger, calm, sad, neutral, happy, disgusted, surprised, fear) and two different intensity levels. We use Ravdess as an emotion recognition evaluation dataset. The SPC (Speech Commands) dataset [212] contains 64,727 total utterances of 30 different words by 1,881 speakers. Table 4.1 summarizes the dataset statistics.

4.3.2 Baselines

In this section, we introduce the other self supervised baseline methods that we compare our proposed method with. We chose a variety of methods that are both audio-only and audio-visual and have varying pretext tasks. When available, we use the original authors' code to evaluate the method. Note that none of these baselines require any labeled data whatsoever.

AVENet [11] is a two-stream audio-visual correspondence based network. One second of audio along with the middle frame of the one second segment are passed as input to the parallel streams, with a positive pair coming from the correct point in the video and a negative pair coming from a different video. The optimization is done with a contrastive loss. We use the audio stream of the network for feature extraction. Korbar et. al. [103] propose an audio-visual temporal synchronization network (Cooperative) which is also a two-stream audiovisual network but has 1 second of video frames as input as opposed to a single frame in AVENet. A positive pair of audio and video samples is one that is in sync, and there are various types of out of sync negative examples in progressive order of difficulty which are optimized with a curriculum learning strategy (easy first, hard later).

Contrast Predictive Coding (CPC) [140] is a technique that tries to model a density ratio to maximize mutual information (MI) between the target signal (random raw audio window) and the context (current raw audio window). By maximizing the MI, the method extracts underlying latent variables that the two different windows have in common. Autoregressive Predictive Coding (APC) [45] is similar to CPC, the key difference being that APC directly tries to predict the future part of the signal based on the history whereas CPC tries to maximize mutual information between the target (future) and the context (present). The input features for APC are 80 dimensional log mel spectrograms with a window size of 25 ms and a step size of 10 ms. The model tries to predict the log mel spectrograms for the future windows given the history. PASE [146] is a self supervised audio encoder trained to predict various features and properties from raw audio. While predicting these multiple attributes (MFCCs, LIM, prosody etc.), the encoder learns a robust, multi-task representation for raw audio that these tasks exemplify (e.g. prosody for emotion).

4.3.3 Evaluation

We evaluate all features on: (i) Discrete Emotion Recognition and (ii) Automatic Speech Recognition.

For the **emotion recognition** task, to investigate how the quality of the representations varies with different types and quantities of training data, we first perform self supervised pretraining on either: i) CREMA-D, a small but emotionally rich database, ii) TCD TIMIT, a medium sized audiovisual speech database or iii) LRW, a large audiovisual speech database. We then use these pretrained models as feature extractors on either CREMA-D or Ravdess to get features for each method. Finally, we train a simple 2 layer LSTM on these features with the hidden size being 256. The learning rate is 0.001 and is decayed by a factor of 0.1 every 30 epochs. We train for 100 epochs and use the weights from the epoch with the best validation accuracy for evaluation. We pass the last hidden state of the LSTM to a linear layer with a size equal to the number of target classes (6 for CREMA, 8 for Ravdess) before a softmax layer with a cross entropy loss for emotion classification. This exact same process (unsupervised feature extraction + LSTM training) is performed for all the methods being compared.

For the **speech recognition** task, we choose the GRID and SPC datasets to evaluate the features. We perform self supervised pretraining for all methods on LRW, and then we use the extracted features converted to Kaldi format for ASR.

Discrete Emotion Recog.		Method (Accuracy $\uparrow$)			
Pretrain Set	Eval Set	AVENet [11]	Co-op [103]	MFCC*	CPC [140]
CREMA-D	CREMA-D	37.72	28.08	41.50	43.61
TCD TIMIT	CREMA-D	28.69	28.81	41.50	30.60
LRW	CREMA-D	29.39	28.53	41.50	34.31
CREMA-D	Ravdess	26.97	17.53	28.32	26.17
TCD TIMIT	Ravdess	20.03	20.72	28.32	26.01
LRW	Ravdess	19.07	21.38	28.32	29.05
Discrete Emotion Recog. (contd.)		Method (Accuracy $\uparrow$)			
Pretrain Set	Eval Set	APC [45]	PASE [146]	Ours	
CREMA-D	CREMA-D	46.80	47.80	55.01	
TCD TIMIT	CREMA-D	39.20	39.32	49.39	
LRW	CREMA-D	41.30	43.16	47.68	
CREMA-D	Ravdess	28.16	23.35	41.34	
TCD TIMIT	Ravdess	32.21	31.76	44.04	
LRW	Ravdess	34.63	30.05	41.92	

Table 4.2: Discrete emotion recognition results (accuracy) presented on the CREMA-D (6 balanced classes, chance = 16.66) and Ravdess (8 balanced classes, chance = 12.5) datasets. All methods are pretrained on the mentioned datasets and used as feature extractors on the evaluation datasets. *MFCC’s are used only in a supervised way on the evaluation datasets.

We employ the ESPNet [213] toolkit for the end-to-end ASR training. We use a hybrid CTC/attention based ASR model with the default ESPNet parameters with a BLSTM encoder with 320 units and location aware attention. We train the model for 15 epochs. For decoding, we use a beam search with a beam size of 20 and a CTC weight of 0.1.

4.4 Results

For **emotion recognition** (Table 4.2), irrespective of the pretraining and evaluation dataset, our method is the best performing by a significant margin. PASE is the closest competing self supervised method when evaluating on CREMA-D, and APC is the closest method on Ravdess. AVENet and Cooperative are not able to learn equally useful emotion representations, likely due to the synchronization pretext task not being the most appropriate for emotion. Our method is also the only one that outperforms the supervised MFCC baseline in every setting. Compared to other methods, it is able to learn more robust and generalizable emotion

Speech Recognition			Method			
Pretrain	Eval	Metric	MFCC	CPC [140]	PASE [146]	Ours
LRW	GRID	Word Error Rate ($\downarrow$)	4.7	10.2	5.8	11.6
LRW	SPC	Accuracy ($\uparrow$)	91.1	74.4	89.1	83.3

Table 4.3: Automatic speech recognition results presented on the GRID and SPC datasets. Compared self-supervised methods are all raw audio encoders. After pre-training as mentioned, features are input to ESPNet for ASR with a hybrid attention/CTC architecture.

features from a variety of pretraining datasets, showing potential for using exponentially larger datasets for self supervised training. These observations indicate that our features are useful unsupervised emotion representations, likely due to our audio features being driven to capture facial expression information (which is highly correlated with acted emotion) by visual supervision.

For **speech recognition** (Table 4.3), we present results only on methods that are raw audio encoders for fair comparison, because methods like APC that encode MFCC’s or Mel features (by directly taking features instead of raw audio signals as inputs) specifically engineered for ASR are likely to be at an advantage. We use the LRW dataset for pretraining all methods evaluated for ASR. On GRID, we achieve a WER of 11.6, while PASE is the best self supervised method with a WER of 5.8. On SPC, we achieve an accuracy of 83.34 on the test set, which is again inferior to PASE with 89.1. CPC outperforms our method on GRID with a WER of 10.2, but is much worse on SPC with an accuracy of 74.37. The supervised baseline is regardless the best performing method for ASR, both on GRID (WER 4.7) and SPC (Accuracy 91.06).

4.5 Discussion

In this chapter, we introduced visually guided self-supervised learning of audio representations of speech using speech-driven facial reconstruction as a pretext task. This was an exploratory work which we further investigate in much more detail (with refined architectures and parameters) in Chapter 5. Visual self-supervision seems to be a promising direction for learning better audio features. Our proposed features significantly outperform existing methods on speech emotion recognition, but there remains a performance gap for speech recognition. We aim to bridge this gap by both making better design choices and also combining the two methods for audio and visual self-supervision that we introduced in Chapters 3 and 4. We

also note that while we used audiovisual speech data in which we reconstruct the video pixel information, there may be other modalities that can offer even better self-supervision (e.g. reconstructing 3D face meshes from audio which is also an active and relevant line of work).

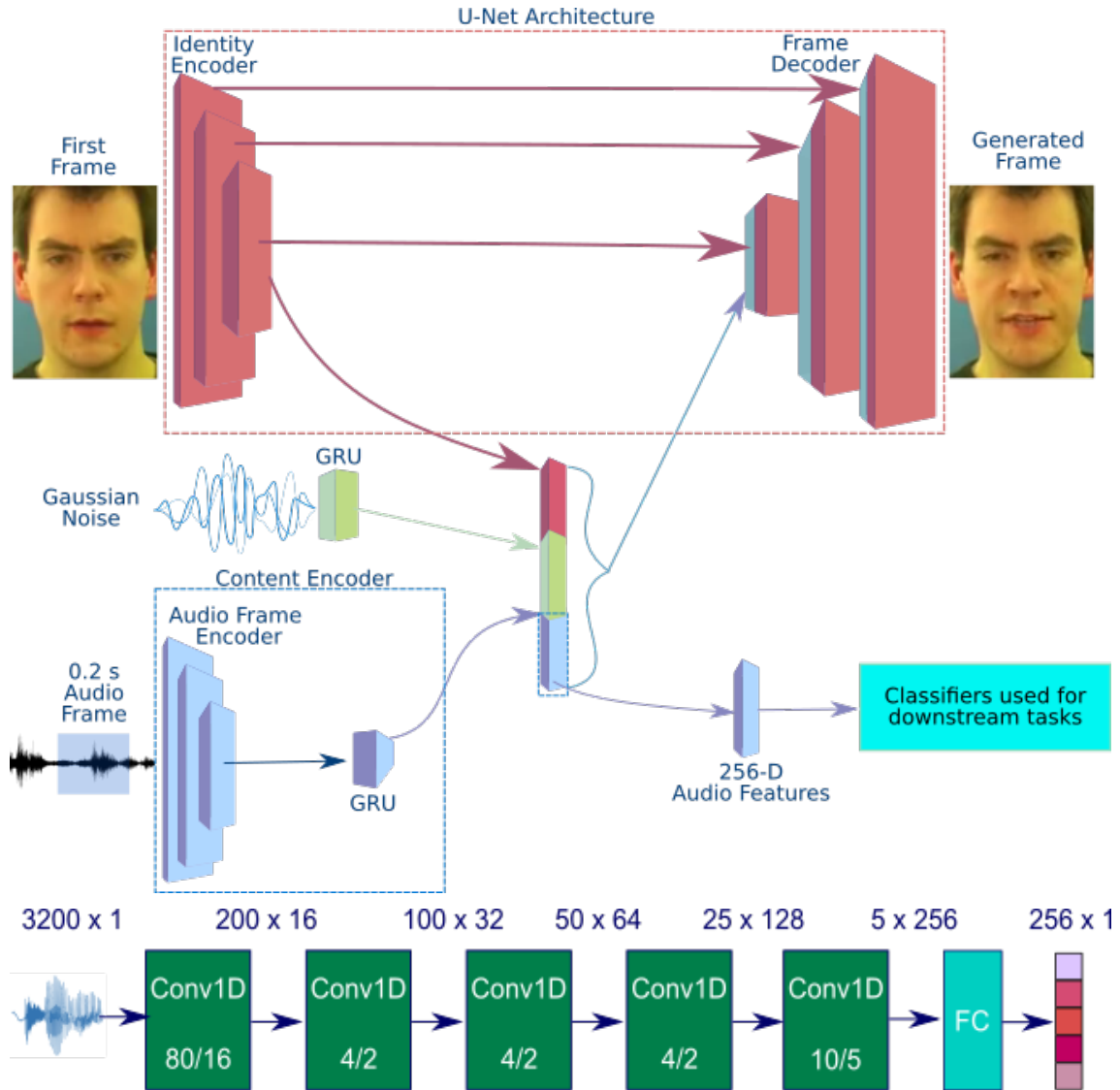


Figure 4.2: (Top) The architecture of the proposed model, (Bottom) the architecture of the audio encoder component which we extract features from after self supervised training.

Chapter 5

Joint Audiovisual Self-Supervision for Speech Representation Learning from Log Mel Inputs

The contents of this chapter are based on the publication

“Does Visual Self-Supervision Improve Learning of Speech Representations for Emotion Recognition?” in IEEE Transactions on Affective Computing, 2021 [188]

In this chapter, we improve upon the methods presented in the previous two chapters. We combine the proposed audio-only and video-only methods for self-supervised learning of speech representations. We find that a combination of audio and visual self-supervision is optimal for the tested datasets.

In this chapter, we investigate joint audiovisual self-supervised learning for audio representations from log mel inputs. Audio representations are a cornerstone of speech and affect recognition. Most affective computing applications involve the analysis of a speech signal using either handcrafted low level descriptors or through a supervised (or finetuned) neural network which directly predicts the labels of interest. However, self-supervised learning may offer better representations for these applications, especially in cases where labeled data is hard to come by and unlabeled audio data is readily available. We look into how self-supervision can be used to produce robust audio features that contain emotional information.

We then propose a novel visual self-supervised method to learn audio features and combine it with the audio-only self-supervised method introduced in Chapter 3. We also show how visual self-supervision helps encode emotional information into the audio features.

Most existing self-supervised learning approaches are unimodal. The few exist-

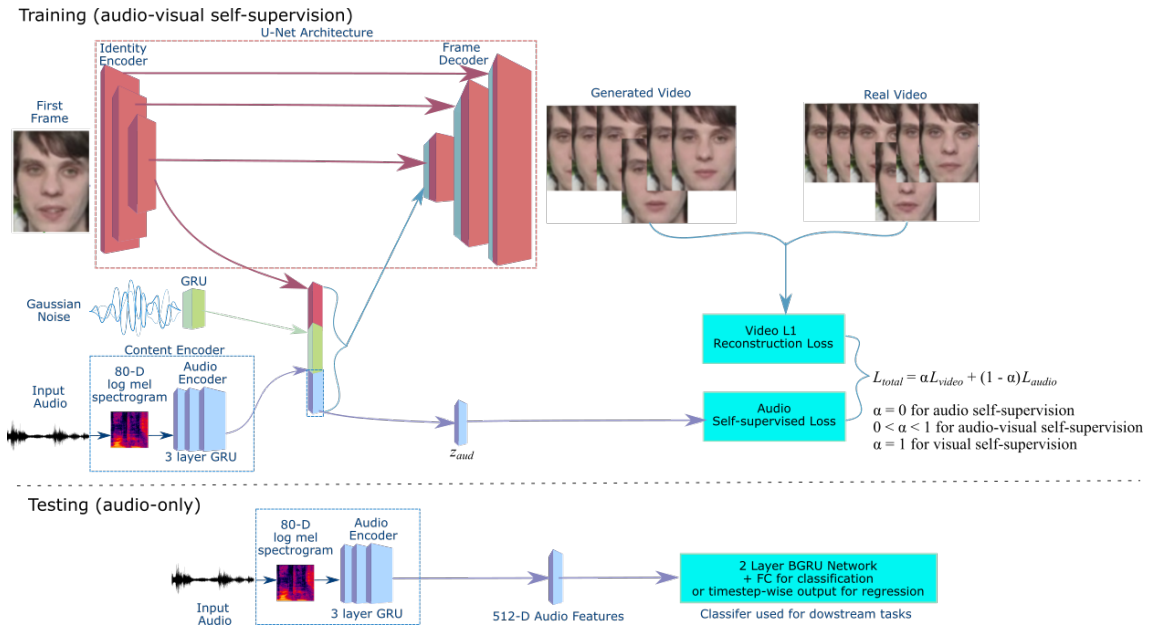


Figure 5.1: An overview of our proposed model for visually guided self-supervised audio representation learning. During training, we generate a video from a still face image and the corresponding audio and optimize the reconstruction loss. An optional audio self-supervised loss can be added to the total to enable multi-modal self-supervision. During testing, we use the audio encoder to extract features for (or finetune on) downstream audio-only tasks.

ing cross-modal approaches typically have some interaction between the modalities in the latent space by pretext tasks like clustering [8] but they do not produce an intuitive interaction between the two modalities. By contrast, our work proposes audio features that are explicitly guided by lip movements and facial expressions' reconstruction (see Fig. 5.1). We implicitly capture visual information related to lip movements and facial expressions in the audio features (see section 5.2.2). The visual modality is needed only during training and our audio features can be evaluated on audio-only datasets.

Contributions

We summarize this chapter's contributions as follows:

1. We investigate visual self-supervision for learning audio features. We propose a novel method for visually-guided self-supervised learning of speech representations by face reconstruction (L1). The proposed speech features, which are correlated with **lip movements** and **facial expressions** due to being driven by video generation, outperform existing audio-only self-supervision approaches for speech and emotion recognition.

2. We propose an audio-only self-supervised method (Odd One Out). Note that this method has been introduced and discussed in detail in Chapter 3. This method is inspired from visual self-supervised learning and is based on temporal order verification as the pretext task. It offers competitive performance on the tested datasets.
3. We combine the proposed audio-only and video-only supervision methods by multi-task learning. We find that the encoder trained in a multi-modal regime encodes richer information about the speech signal and yields the most effective representation that attains the best performance among all tested methods.
4. We show that pretraining by audio-visual self-supervision produces a better weight initialization for downstream tasks than does training from scratch. This is especially relevant in the context of affective computing due to reduced overfitting and improved performance on small datasets.
5. We show that the proposed visually-guided audio features are more robust for various levels of noise.

5.1 Background

Related work for this chapter has been introduced in detail in Chapter 2

5.2 Combining Audio and Visual Self-Supervision

5.2.1 Visual Self-Supervision for Speech Representation Learning

The proposed method is illustrated in Fig. 5.1 and is based on our prior work on visually guided speech representation learning through speech-driven facial animation [206, 187, 185, 186]. The model is a temporal encoder-decoder which takes a still image of a face (frame from a 25 fps video) and an audio signal as inputs and generates video frames from these. The model itself can be conceptually divided into three subnetworks (see Fig. 5.1 and Fig. 5.2), namely the content/audio encoder (3 layer GRU), the identity encoder (6 layer 2D CNN) and the frame decoder (U-Net [170] architecture with skip connections from the identity encoder).

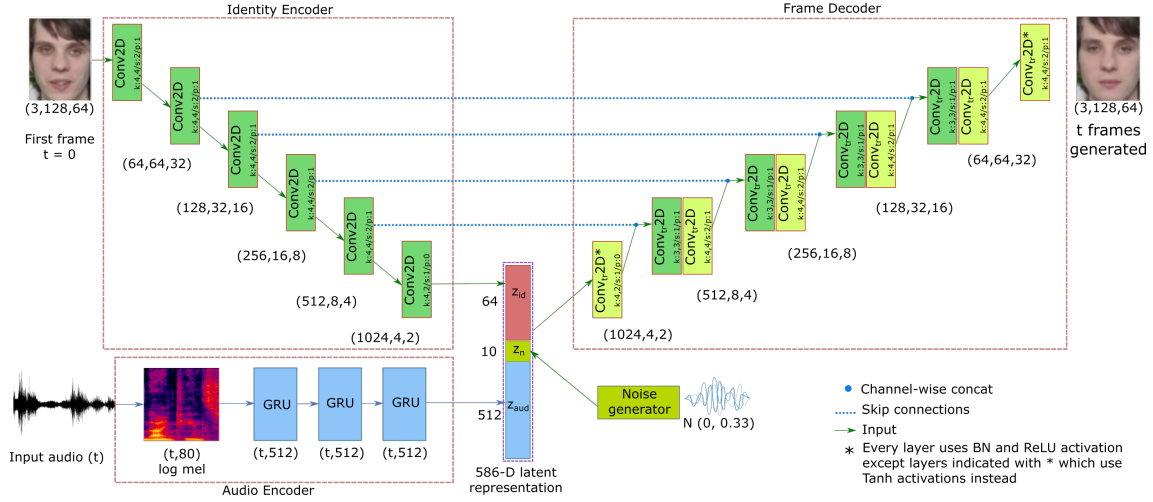


Figure 5.2: A detailed illustration of our the encoder-decoder model we use for video reconstruction. From an unlabeled sample of audiovisual speech, we use the audio and the first frame of the video ($t = 0$) to generate a video with t frames. The model contains: (1) an identity encoder which produces a 64-D identity embedding; (2) an audio encoder which converts the input audio (t frames of 80 dimensional log mel spectrograms) into a 512-D audio embedding; (3) a frame decoder which generates video from the concatenated latent representation using transposed convolutions.

The architecture of the content encoder is a 3 layer GRU with log mel spectrograms as input (closely following [45]), as shown in Fig. 5.2. Its purpose is to convert the input audio into a latent space audio feature vector z_{aud} . Similarly, the identity encoder (see Fig. 5.2 top-left), which is made of 6 (Conv2D - BatchNorm - ReLU) blocks, reduces a 64x128 input image (which is the first video frame of the audiovisual speech segment) to a 64x1 feature vector z_{id} .

We also use a noise generator (see Fig. 5.1) capable of producing noise that is temporally coherent. A 10 dimensional vector is sampled from a Gaussian distribution with mean 0 and variance of 0.33 and passed through a single-layer GRU to produce the noise sequence. This latent representation z_n accounts for randomness in the face synthesis process (such as the generation of random sequential behaviour like blinks [208]), which leads to a more realistic facial reconstruction.

The latent representation is the concatenation of z_{aud} , z_{id} and z_n (as shown in Fig. 5.2). This results in a 586 dimensional embedding. This embedding then goes through the frame decoder (see Fig. 5.2 top-right), which is a CNN that uses strided transposed convolutions to produce the video frames. The skip connections to the identity encoder help in preserving subject identity.

An L1 reconstruction loss between a random frame from the generated video and the corresponding frame from the real video is used to train the network. The L1 loss on the pixel level is commonly used in facial reconstruction as opposed to

the L2 loss which typically produces blurrier reconstructions. We use the Adam optimizer with a learning rate of 0.06 that is decayed by a factor of 0.98 every 10 epochs. Essentially, our model aims to predict the video modality (full face reconstruction) given only the audio modality and speaker identity information from the first frame. In this process, the audio encoder is driven to produce useful **speech features that correlate with mouth and facial movements** (because we need to generate these lip and facial movements using only the audio information, so the features z_{aud} must encode this in order to reduce the L1 loss). After this process of visually guided self-supervised pretraining, we simply use the trained audio encoder as a pretrained model for audio-only downstream tasks. The features extracted from this model are especially interesting to evaluate on tasks like speech recognition and emotion recognition. This is because these features are explicitly trained (guided by the visual modality) to contain information related to lip movements (highly correlated with speech content) and facial expressions (highly correlated with emotion).

5.2.2 Why does Visual Self-Supervision work for Emotion?

To support our claim that the proposed method produces good features for emotion, we present qualitative results for full face reconstruction using our method. We use the same neutral expression still frame from the CREMA-D dataset as the input image and produce video sequences driven by different audio clips of multiple emotions (see Fig. 5.3). We observe that facial expressions that correlate with the emotion of the audio clip are generated in the output video. Note that this is achieved using only a neutral still image and the audio information. This offers evidence that emotion related information is being encoded in the audio features being used to drive the generation. Our method is able to generate a variety of facial expressions that correlate with the emotion of the input audio. These observations form the basis of our hypothesis that the proposed task (reconstructing the full face based only on the audio information) can help produce better audio features for emotion recognition in a self-supervised manner. The reader can find more detailed insights into the speech-driven facial animation process in our prior work [206, 208].

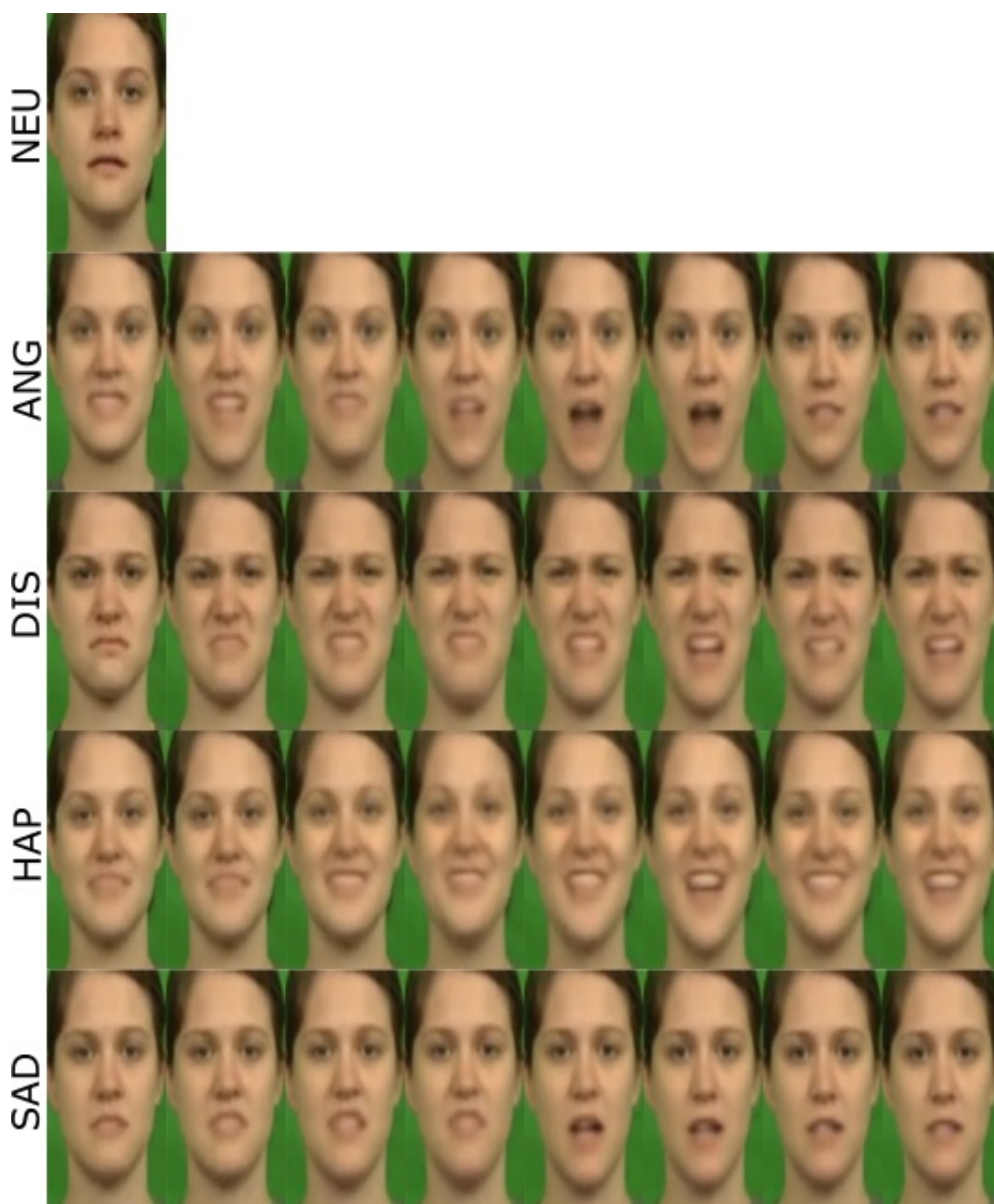


Figure 5.3: Videos produced by the proposed method using the same neutral still frame taken from the CREMA-D test set and driven by the sentence “its eleven o’clock” spoken with multiple emotions (Anger, Disgust, Happiness, Sadness). This highlights that emotional information is captured in the audio features being used for the generation.

5.2.3 Audio Encoder Architecture

The audio encoder (see Fig. 5.2 bottom-left) is a log mel spectrogram encoder (closely following [45]). The log mel spectrogram is computed with 80 frequency bins, a window width of 25ms and a stride of 10ms, which is a standard choice for processing speech signals. This (t, 80) dimensional input then goes through the encoder which is a 3 layer GRU network with each layer having a hidden size of 512 followed by a fully connected layer which converts it into a feature with dimensionality (t, 512). This specific architecture of the audio encoder with 3 GRU layers is the exact same as used in [45]. We chose this for simplicity and to enable direct comparison with this baseline which uses a similar audio input as us (80 dimensional log mel features).

We use the above described architecture (see Fig. 5.2) as the audio encoder in the proposed models in the following sections (for both visual and audio self-supervision). We then use the trained encoder to extract features from the evaluation datasets. We implement our method and all evaluation pipelines using PyTorch [147].

5.2.4 Audiovisual Self-Supervision for Speech Representation Learning

We combine the proposed audio and visual self-supervision methods by making the encoder jointly predict the visual self-supervision task and the audio self-supervision task. Since we used the same encoder architecture for both the visual and audio tasks, this is straightforward to accomplish. In the pipeline shown in Fig. 5.1 for visual self-supervision, we also use the optional prediction for the audio-only self-supervised task (Odd). This leads to two losses being calculated, one for visual and one for audio self-supervision.

The total loss L_{total} is the weighted sum of the L1 reconstruction loss from visual self-supervision L_{video} and the cross entropy loss from the audio-only self-supervision L_{audio} . α is the weight factor which controls how much of the loss term comes from which type of supervision. The total loss is given by the equation:

$$L_{total} = \alpha L_{video} + (1 - \alpha) L_{audio} \quad (5.1)$$

5.3 Experimental Setup

5.3.1 Datasets

This section introduces the various audio-only and audiovisual datasets that were used in the work either for pretraining or evaluating the baseline and proposed models. For all datasets, we divide the data into training, validation and test sets with all samples from each speaker belonging to a particular set only. Table 5.1 summarizes the statistics for all the datasets used in this work.

The CREMA-D dataset [25] contains a diverse set of 91 actors who utter 12 sentences multiples times each with a different level of intensity for each of 6 basic emotional labels (anger, fear, disgust, neutral, happy, sad). We use Crema as a discrete emotion recognition evaluation dataset.

The Ravdess dataset [115] contains 1440 samples of 24 different actors who acted out two sentences with 8 different basic emotions (anger, calm, sad, neutral, happy, disgusted, surprised, fear) and two different intensity levels. We use Ravdess also as a discrete emotion recognition evaluation dataset.

The RECOLA dataset [168] contains dyadic conversations in French between a pair of participants working to solve a collaborative task over video conference. The annotated part of the dataset consists of 5-minute long clips (with audio from one speaker only) with continuous valence and arousal annotations from 6 annotators. Since the labels of the test set are not publicly available, we report the performance of our models on the development set.

The SEWA dataset [104] contains dyadic conversations over video conference between a pair of participants discussing about an advertisement that they have just watched. The audio clips are typically 3-minute long and have continuous valence and arousal annotations. We use the AVEC2017 challenge German language subset (with the gold standard valence and arousal labels) of the dataset to run our experiments.

The IEMOCAP dataset [24] contains dyadic conversations between 10 speakers for a total of 12 hours of audiovisual data. The discrete emotion labels comprise of 8 categories (anger, happiness, sadness, neutral, excitement, frustration, fear, surprise), however we only consider the first 4 categories for our experiments (anger, happiness, sadness, neutral). This is due to much higher inter annotator agreement for these categories, and this portion of the dataset has been similarly used in prior studies [122]. This partition also leaves us with around 6.5 hours of data instead of the original 12 hours. We use IEMOCAP as another discrete

emotion recognition evaluation dataset.

The SPC (Speech Commands) dataset [212] contains 64,727 total utterances of 30 different words by 1,881 speakers. We use SPC as a speech recognition evaluation dataset.

The LRW dataset [41] is a large, in-the-wild dataset of 500 different isolated words primarily from BBC recordings. It is an audiovisual speech dataset and is thus appropriate for training our methods. We use a subset of LRW that has only nearly frontal videos (with yaw, pitch and roll restricted to a maximum of 10 degrees), in order to have a cleaner supervisory signal from the visual modality. This filtering leaves us with a total of around 40 hours of usable data. We use LRW as the self-supervised pretraining dataset for all baseline and proposed methods.

5.3.2 Baselines

Since the aim of our work is to yield self-supervised audio features, we compare against other baselines focusing on the same goal. The three self-supervised methods we compare against are CPC [140], APC [45] and PASE [146]. In our prior work [187], we have also used other audio-visual self-supervised methods as baselines [11, 103], but due to them not obtaining competitive performance on the target datasets (CREMA/SPC), we exclude them from this paper.

Contrast Predictive Coding (CPC) [140] is a technique that tries to model a density ratio to maximize mutual information (MI) between the target signal (random raw audio window) and the context (current raw audio window). By maximizing the MI, the method can extract the underlying latent variables that the two different parts of the signal have in common. This essentially means that the representations of related (positive) examples lie closer in the latent space, and those of negative examples lie further away.

Autoregressive Predictive Coding (APC) [45] is similar to CPC, however the key difference is that APC directly tries to predict the immediate future part of the signal based on the history whereas CPC tries to maximize mutual information between the target (future) and the context (present). The input features for APC are 80 dimensional log mel spectrograms with a window size of 25 ms and a step size of 10 ms. The model tries to predict the log mel spectrograms for the future windows given the history.

PASE [146] is a raw audio encoder trained in a self supervised way to predict various different handcrafted features such as MFCC, prosody, waveform etc. While predicting these multiple tasks, the encoder learns a very robust and multi-

Dataset	Train	Val	Test
LRW	112658 / 36.3	5870 / 1.90	5980 / 1.90
SPC	51094 / 14.2	6798 / 1.88	6835 / 1.89
CREMA-D	11594 / 9.70	819 / 0.70	820 / 0.68
Ravdess	1509 / 1.76	415 / 0.48	519 / 0.60
IEMOCAP	3548 / 4.28	793 / 0.95	942 / 1.31
RECOLA	180 / 1.00	150 / 0.80	-
SEWA	706 / 0.78	287 / 0.31	352 / 0.39

Table 5.1: The number of samples and duration (number / time in hours) of speech data in the training, validation and test sets of each dataset.

task representation for raw audio that these tasks exemplify (e.g. prosody for emotion).

We also compare our methods against 39 dimensional MFCCs (13 coefficients, 13 deltas, and 13 delta-deltas) which act as baseline features used for supervised learning for audio. We also compare with 80 dimensional log mel spectrograms. Additionally, we also use the OpenSmile [59] toolkit to extract LLD’s (low level descriptors) based on the IS13-ComParE feature set. These features are well established baselines for audio based affective computing tasks.

5.4 Results

This section presents the details of all experiments that we perform to rigorously validate our proposed method. We present all results for discrete emotion recognition (Table 5.2), continuous affect recognition (Table 5.3), and speech recognition (Table 5.4) from the extracted features. We present results for both visual and audio self-supervision for all variants of the models. We also show the results with the combination of the visual and audio self-supervision approaches using multi-task learning. We present numerous ablation studies such as the variation of model performance with change in pretraining set size and noise level. We also compare the frozen encoders (that provide fixed extracted features) with their finetuned (pretrained weights are also updated in downstream training) and fully supervised (trained from scratch on the downstream task) equivalents.

5.4.1 Experimental Setup

We evaluate all extracted features on: (i) Discrete Emotion Recognition, (ii) Continuous Affect Recognition, and (iii) Automatic Speech Recognition (ASR).

Downstream Results		Discrete Emotion		
Pretraining Dataset	Evaluation Dataset	LRW CREMA-D	LRW Ravdess	LRW IEMOCAP
Labels		6 emotions	8 emotions	4 emotions
Method	Pretext task	F1 Score ($\uparrow$)	F1 Score ($\uparrow$)	F1 Score ($\uparrow$)
MFCC	None	0.447 ± 0.012	0.390 ± 0.013	0.535 ± 0.006
Log Mel	None	0.463 ± 0.013	0.480 ± 0.010	0.572 ± 0.011
LLD [59]	None	0.409 ± 0.015	0.398 ± 0.013	0.497 ± 0.013
CPC [140]	Audio	0.422 ± 0.016	0.408 ± 0.011	0.517 ± 0.011
PASE [146]	Audio	0.488 ± 0.010	0.472 ± 0.010	0.557 ± 0.017
APC [45]	Audio	0.480 ± 0.012	0.497 ± 0.011	0.592 ± 0.019
Odd	Audio	0.542 ± 0.016	0.501 ± 0.008	0.615 ± 0.010
L1	Visual	0.567 ± 0.014	0.560 ± 0.010	0.618 ± 0.009
L1 + Odd	Audiovisual	0.573 ± 0.012	0.565 ± 0.008	0.631 ± 0.007
L1 + Odd (finetuned)	Audiovisual	0.592 ± 0.010	0.645 ± 0.016	0.642 ± 0.011
Supervised	None	0.540 ± 0.013	0.528 ± 0.010	0.623 ± 0.010

Table 5.2: Results for all baseline and proposed methods for discrete emotion recognition (on CREMA, Ravdess, and IEMOCAP). Unless specified otherwise, all methods are used as frozen audio feature extractors before training a classifier (BGRU) on the downstream task. All results are in the form of $\mu \pm \sigma$ for 10 runs.

Downstream Results		Continuous Affect Recognition	
Pretraining Dataset		LRW	
Evaluation Dataset		SEWA	
Labels		Valence	Arousal
Method	Pretext task	CCC ($\uparrow$)	CCC ($\uparrow$)
MFCC	None	0.349 ± 0.009	0.359 ± 0.010
Log Mel	None	0.307 ± 0.011	0.316 ± 0.007
LLD [59]	None	0.245 ± 0.011	0.236 ± 0.010
CPC [140]	Audio	0.287 ± 0.008	0.295 ± 0.009
PASE [146]	Audio	0.354 ± 0.008	0.343 ± 0.010
APC [45]	Audio	0.317 ± 0.009	0.321 ± 0.014
Odd	Audio	0.342 ± 0.008	0.339 ± 0.010
L1	Visual	0.361 ± 0.008	0.368 ± 0.009
L1 + Odd	Audiovisual	0.373 ± 0.007	0.371 ± 0.009
L1 + Odd (finetuned)	Audiovisual	0.380 ± 0.010	0.383 ± 0.009
Supervised	None	0.297 ± 0.019	0.330 ± 0.009

Downstream Results (contd.)		Continuous Affect Recognition	
Pretraining Dataset		LRW	
Evaluation Dataset		RECOLA	
Labels		Valence	Arousal
Method	Pretext task	CCC ($\uparrow$)	CCC ($\uparrow$)
MFCC	None	0.413 ± 0.008	0.714 ± 0.011
Log Mel	None	0.396 ± 0.012	0.696 ± 0.011
LLD [59]	None	0.271 ± 0.015	0.624 ± 0.010
CPC [140]	Audio	0.320 ± 0.011	0.623 ± 0.012
PASE [146]	Audio	0.390 ± 0.008	0.719 ± 0.007
APC [45]	Audio	0.383 ± 0.010	0.703 ± 0.008
Odd	Audio	0.436 ± 0.012	0.720 ± 0.008
L1	Visual	0.438 ± 0.011	0.733 ± 0.012
L1 + Odd	Audiovisual	0.448 ± 0.009	0.745 ± 0.008
L1 + Odd (finetuned)	Audiovisual	0.452 ± 0.012	0.764 ± 0.006
Supervised	None	0.402 ± 0.010	0.701 ± 0.008

Table 5.3: Results for all baseline and proposed methods for continuous affect recognition (on SEWA and RECOLA). Unless specified otherwise, all methods are used as frozen audio feature extractors before training a classifier (BGRU) on the downstream task. All results are in the form of $\mu \pm \sigma$ for 10 runs.

Results		Speech Recognition
Pretraining Dataset		LRW
Evaluation Dataset		SPC
Labels		30 words
Method	Pretext task	F1 Score ($\uparrow$)
MFCC	None	0.912 ± 0.002
Log Mel	None	0.927 ± 0.003
LLD [59]	None	0.643 ± 0.014
CPC [140]	Audio	0.814 ± 0.019
PASE [146]	Audio	0.941 ± 0.002
APC [45]	Audio	0.933 ± 0.003
Odd	Audio	0.919 ± 0.002
L1	Visual	0.935 ± 0.002
L1 + Odd	Audiovisual	0.944 ± 0.002
L1 + Odd (finetuned)	Audiovisual	0.953 ± 0.002
Supervised	None	0.921 ± 0.005

Table 5.4: Results for all baseline and proposed methods for speech recognition (on SPC). Unless specified otherwise, all methods are used as frozen audio feature extractors before training a classifier (BGRU) on the downstream task. All results are in the form of $\mu \pm \sigma$ for 10 runs.

For the **emotion recognition** task, we first perform self-supervised pretraining on LRW as described, and then use the pretrained models as feature extractors on the CREMA, Ravdess and IEMOCAP datasets. Once we have these features, we then train an BGRU (bidirectional gated recurrent unit) model for the emotion classification task. We used a BGRU for simplicity, however this can be replaced by any model that can classify variable length sequences into discrete categories (such as LSTMs, TCNs, Transformers, LiGRUs [166]). We use a 2 layer BGRU with 256 units in each layer. The initial learning rate is 0.0001 and is decayed by a factor of 0.1 every 40 epochs. We train for 100 epochs and use the checkpoint from the epoch which gives the best validation set metric for evaluation on the test set. We pass the last hidden state of the BGRU to a linear layer with size equal to the number of target classes (6 for CREMA, 8 for Ravdess, 4 for IEMOCAP) followed by a Softmax layer with a cross entropy loss for emotion classification. This process (self-supervised feature extraction + BGRU training) is followed for all compared methods (see bottom of Fig. 5.1).

For the **speech recognition** task, we use the SPC dataset to evaluate our method. For the SPC dataset which is a spoken word classification task with 30 different possible labels, we use the exact same protocol as described for emotion recognition (self-supervised feature extraction + BGRU training). We use the same parameters and learning schedule for the BGRU.

For the **continuous affect recognition** task, we use SEWA and RECOLA as the evaluation datasets. We predict continuous values of valence and arousal for each audio segment. We train separate individual models for valence and arousal prediction. Using the same encoder architecture and self-supervised pretraining followed by downstream evaluation process described in the bottom of Fig. 5.1, we use the output of the BGRU at each timestep as the predicted valence or arousal value. This gives us a predicted sequence y . We train the network to maximize the Concordance Correlation Coefficient (CCC) between the predicted sequence y and the ground truth valence or arousal sequence $\hat{y}$. The formula for CCC is:

$$CCC = \frac{2\sigma_{y\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2} \quad (5.2)$$

where μ_y and $\mu_{\hat{y}}$ are the means of the ground truth and predicted sequences, σ_y^2 and $\sigma_{\hat{y}}^2$ are the variance of the ground truth and predicted sequences, and $2\sigma_{y\hat{y}}$ is the covariance between the ground truth and predicted sequence. The value of CCC lies between -1 and 1, and is 0 for uncorrelated sequences. The loss function that we minimize to train our network (which maximizes CCC) is 1 -

$((CCC+1)/2)$, which lies between 0 and 1. For SEWA, we report the CCC on the test set as the evaluation metric. For RECOLA, we report the CCC on the validation set to because test set labels are not publicly available. Note that CCC is the standard metric for the SEWA dataset (used in the AVEC 2017 challenge) and the RECOLA dataset (used in the AVEC 2016 challenge).

For all results in Tables 5.2, 5.3, and 5.4, we validate the models using the validation set of the downstream task and use the model from the epoch with the best validation performance to evaluate on the test set. We report the results in the form of $\mu \pm \sigma$ (mean $\pm$ std) over 10 runs for each experiment.

5.4.2 Results with Visual Self-Supervision (L1)

Our method for visual self-supervision by face reconstruction from audio is based on an L1 reconstruction loss, and is indicated as **L1** throughout the results in Tables 5.2, 5.3, and 5.4. We also use a pairwise t-test between observations over 10 runs when discussing the statistical significance of the results throughout Section 5.4. For discrete emotion recognition, irrespective of dataset, our method performs better than any audio self-supervised baseline. On CREMA, L1 achieves an F1 score of 0.567, which is significantly better ($p < 0.01$) than the best performing baseline PASE, which achieves an F1 score of 0.488. For Ravdess, APC is the best baseline with an F1 score of 0.497, but L1 with an F1 score of 0.560 significantly outperforms this ($p < 0.01$). The same trend can be seen for IEMOCAP, with L1 again being the best performing method with an F1 score of 0.618. For speech recognition on SPC, L1 is again the best self-supervised method with an F1 score of 0.935, which outperforms MFCCs (optimised for speech recognition) at 0.912 (significant difference, $p < 0.01$). For continuous affect recognition on SEWA, L1 outperforms the other self-supervised methods with a CCC of 0.361 for valence and 0.378 for arousal. The closest baselines to L1 on SEWA are PASE for valence (CCC=0.354, $p=0.136$) and MFCCs for arousal (CCC=0.359, $p=0.051$). On RECOLA, L1 again outperforms the other methods with a valence CCC of 0.438 and an arousal CCC of 0.733. The closest competing methods to L1 are again PASE (arousal CCC=0.719, $p < 0.01$) and MFCCs (valence CCC=0.413, $p < 0.01$).

In summary, the proposed method for visual self-supervision leads to features that significantly outperform those from baseline audio self-supervised methods for both emotion recognition and speech recognition.

5.4.3 Results with Audio-only Self-Supervision (Odd)

Our method for audio-only self-supervision (Odd One Out) is indicated as **Odd** throughout the discussion of the results. Results for audio-only self-supervision are discussed in this chapter for completeness and context. We include only results for Odd (and not AoT) as it was the better performing proposed method on average across tasks. A more detailed discussion of audio-only results can be found in Chapter 3.

For discrete emotion recognition, Odd is the best performing audio-only self-supervised method on CREMA (F1 score of 0.542), Ravdess (F1 score of 0.501) and IEMOCAP (F1 score of 0.615). For continuous affect recognition on SEWA, Odd achieves a valence CCC of 0.342, which is outperformed by MFCCs with CCC of 0.349 ($p=0.13$). For arousal, Odd achieves CCC of 0.339 which is outperformed by both MFCCs and PASE with CCCs of 0.359 ($p<0.01$) and 0.343 ($p=0.362$) respectively. On RECOLA, Odd is the best method for arousal (CCC=0.720) and valence (CCC=0.436). For speech recognition on SPC, both PASE (F1 score of 0.941) and APC (F1 score of 0.933) significantly ($p<0.01$) outperform Odd (F1 score of 0.919).

We notice that while Odd is the best method seen so far for discrete emotion and offers competitive performance with respect to other methods on the other tasks, there is a performance gap for speech recognition and continuous affect recognition with MFCCs and PASE. We aim to bridge this performance gap and yield better features using a multimodal combination of the proposed methods using multi-task learning.

5.4.4 Results with Audiovisual Self-Supervision (L1 + Odd)

In order to determine the optimal weights for each modality for multi-task learning, we tune the parameter α (equation 5.1) on the validation sets of the CREMA, Ravdess and SPC datasets (introduced in section 5.3.1) for a range of values. The results for the tuning are in Table 5.6. From the table, we observe that the best value of α is 0.67 for L1 + Odd. Thus, we use the model trained with this value when evaluating on the test sets in all experiments.

When comparing the results with other results in Tables 5.2, 5.3, and 5.4, we can see a clear improvement using audiovisual self-supervision. L1 + Odd significantly outperforms all other methods in every experiment. For discrete emotion recognition, L1 + Odd is the best-performing method on CREMA (F1 score = 0.573), IEMOCAP (F1 score = 0.631), and Ravdess (F1 score = 0.565). For

speech recognition, L1 + Odd is the best-performing method on SPC (F1 score = 0.944). The notable observation here is that L1 + Odd significantly outperforms either unimodal method in all experiments ($p < 0.01$ for all differences from Odd, $p < 0.05$ for all differences from L1, except for CREMA with $p = 0.33$ and arousal on SEWA with $p = 0.37$). For speech recognition, this also helps in outperforming PASE (difference with $p < 0.05$) and MFCCs (difference with $p < 0.01$), which Odd was not able to do on its own. This points to the presence of complementary information being encoded by the two types of supervision from the two modalities which leads to very good generalized audio representations. In summary, the proposed multimodal self-supervision method clearly outperform any proposed unimodal self-supervision method.

5.4.5 When does Visual Self-Supervision help specifically?

We now examine the specific cases and scenarios where visual information is especially helpful to improve the audio representations. For each of the three downstream tasks (discrete emotion, continuous affect, and speech recognition), we compare the class-wise predictions of the model trained using only audio self-supervision (Odd) and that with joint audiovisual self-supervision (L1 + Odd).

For discrete emotion recognition (see Fig. 5.4), we analyze the difference in performance between the model trained using only audio self-supervision (Odd) and the model trained using joint audiovisual self-supervision (L1 + Odd) for each emotion class in the CREMA dataset. We saw in Table 5.2 that L1 + Odd gets an overall F1 score of 0.573 compared to Odd with F1 score of 0.542. In Fig. 5.4, we see how each class contributes to this performance gain. We find that the gain is the highest for the HAP (happy) and NEU (neutral) classes. There are also moderate gains seen for the ANG (anger), DIS (disgust) and SAD (sadness) classes. However there is a performance drop observed for the DIS (disgust) class. Overall, we conclude that visual information during pretraining helps reduce the confusion between classes such as HAP and NEU (which are more clearly separable in the visual modality than the audio modality). Even though the DIS class is negatively impacted, this results in a significantly better audio representation for discrete emotion.

For continuous affect recognition (see Table 5.5), we analyze the difference in performance between Odd (audio self-supervision), L1 (visual self-supervision) and L1 + Odd (audiovisual self-supervision) for different types of audio segments in the RECOLA dataset. We compare the methods for four different types of audio

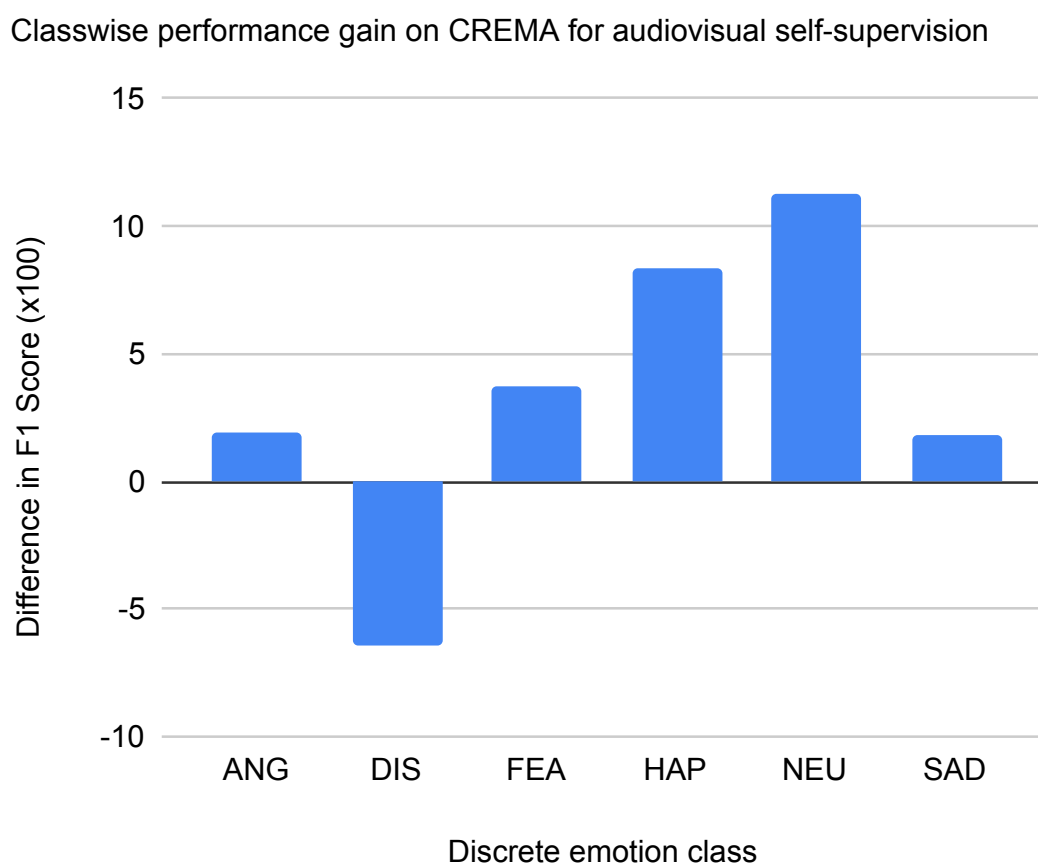


Figure 5.4: The classwise performance gain of L1 + Odd (audiovisual) compared to Odd (audio-only) on the CREMA dataset. Values are 100x the difference in F1 scores.

Relative CCC (x100) gain	Method		
	Audio	Visual	Audiovisual
Low Valence	0.00	+0.02	+0.44
High Valence	0.00	+2.15	+6.76
Low Arousal	0.00	+1.10	+2.87
High Arousal	0.00	+1.01	+2.91

Table 5.5: Comparison of Odd (Audio), L1 (Visual) and L1 + Odd (Audiovisual) on high/low valence and high/low arousal audio segments from the RECOLA dataset. Performance gains are relative to the CCC attained by Odd on each type of segment. Gain values are CCCx100.

segments (low/high valence and low/high arousal). We use the test CCC attained by Odd as the baseline and compare the performance gain attained with respect to this. We observe that L1 + Odd is the best method for all types of segments. The performance gain from visual information is most pronounced for high valence sequences (+2.15 for L1 and +6.76 for L1 + Odd). We see a very slight gain for low valence sequences (+0.02 for L1 and +0.44 for L1 + Odd). We also see overall gains for both high and low arousal (+1.01/+1.10 for L1 and +2.91/+2.87 for L1 + Odd). Overall, L1 + Odd still offers the best representation for both valence and arousal.

For speech recognition (see Fig. 5.5), we analyze the class-wise performance difference in performance on the SPC dataset between Odd (audio self-supervision) and L1 + Odd (audiovisual self-supervision). We present a histogram of difference in class-wise performance in Fig. 5.5. We see that most classes are positively influenced with a gain in the range of 2% to 6%, with 2 of the 30 classes seeing gains of nearly 8%. The classes with the highest improvements due to visual information primarily are similar sounding words such as 'no'/'go', 'tree'/'three', and 'bird'/'bed'. All of these classes have similar audio and thus higher intra-class confusion, but have useful distinguishing information in the video modality which our proposed audiovisual self-supervision is able to exploit. While the performance gain is the most significant for these specific classes, we also see an overall performance gain for every class. Thus, L1 + Odd offers the best representation for speech recognition on SPC.

Classwise performance gain on SPC for audiovisual self-supervision

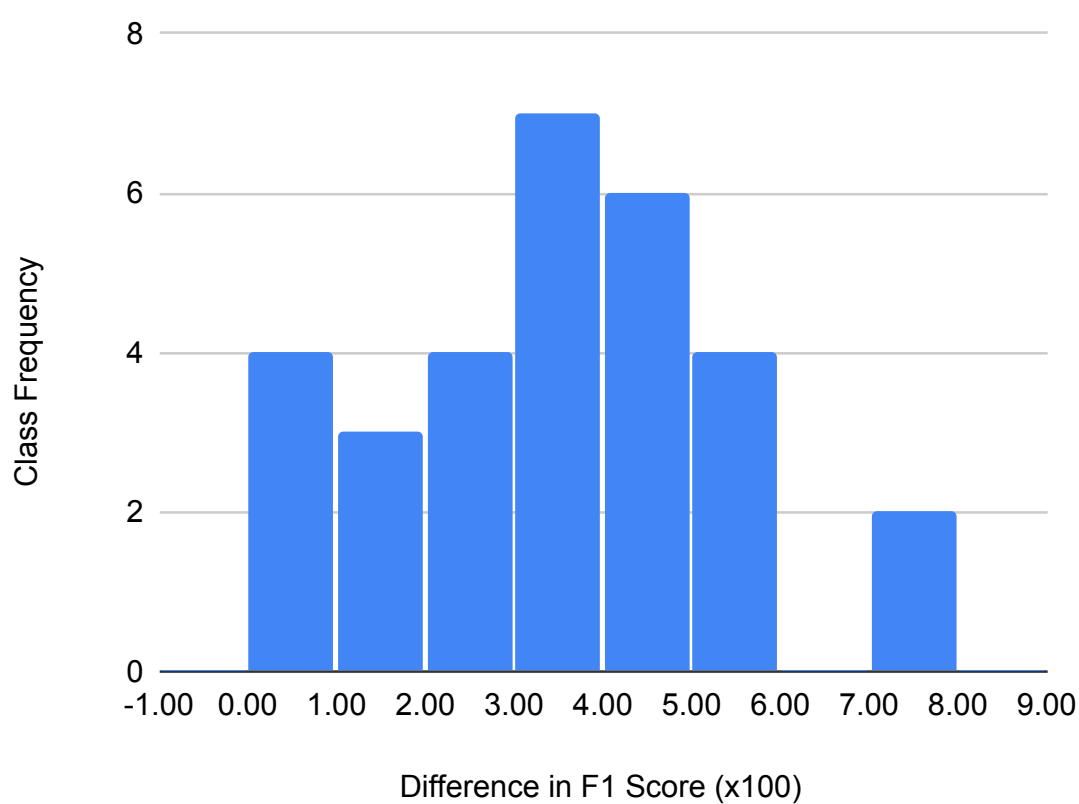
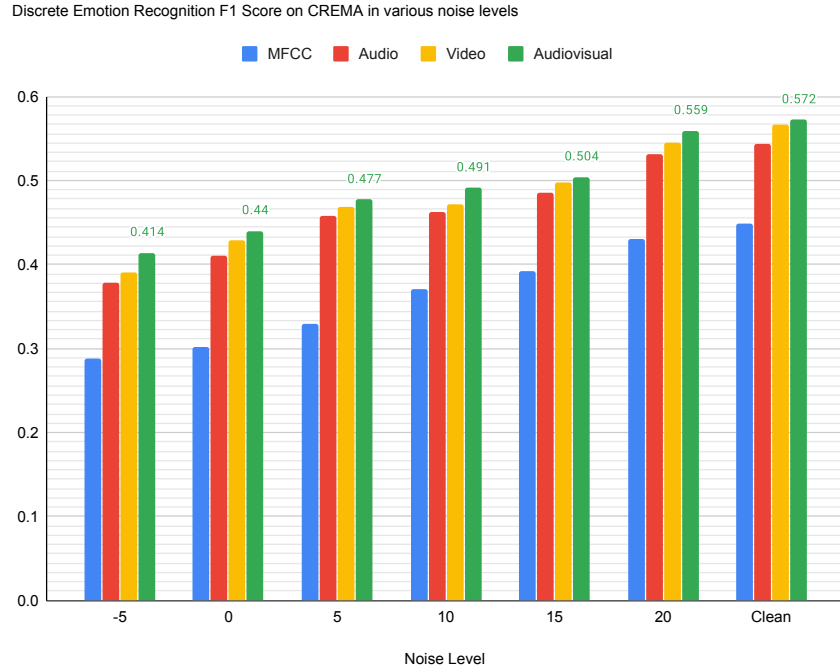


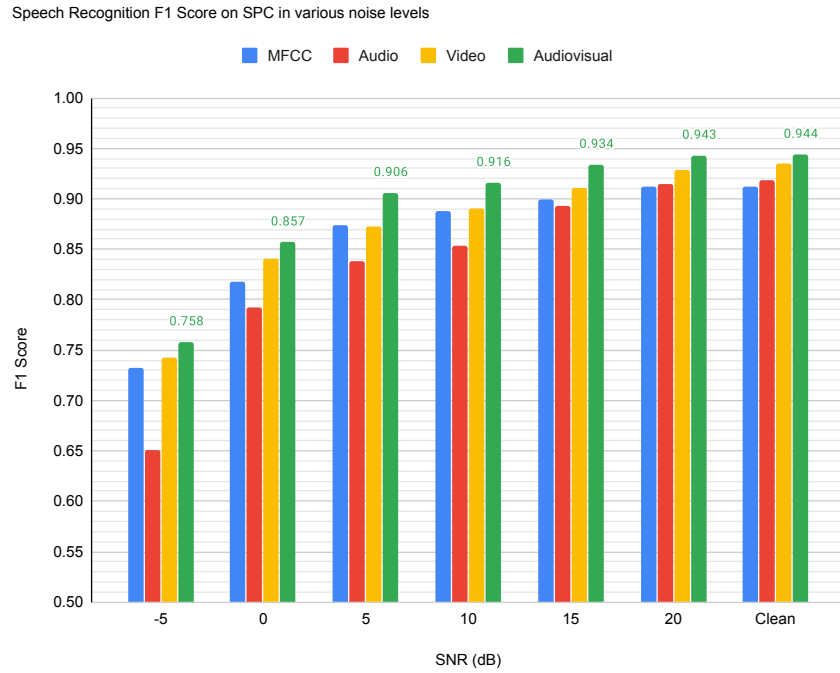
Figure 5.5: Histogram of the classwise performance gain of L1 + Odd (audiovisual) compared to Odd (audio-only) on the SPC dataset. Values are 100x the difference in F1 scores.

MTL Weight Tuning for Audio and Visual Tasks			Emotion Recognition	
	Pretraining Dataset		LRW	LRW
	Evaluation Dataset		CREMA-D	Ravdess
	Classifier for $(t, 512)$ features		LSTM	LSTM
	Labels		6 emotions	8 emotions
Method	Video weight (α)	Audio weight ($1 - \alpha$)	Accuracy ($\uparrow$)	Accuracy ($\uparrow$)
L1 + AoT	0.17	0.83	46.22	38.61
L1 + AoT	0.33	0.67	47.91	40.18
L1 + AoT	0.50	0.50	51.50	43.03
L1 + AoT	0.67	0.33	51.77	44.39
L1 + AoT	0.83	0.17	48.93	40.40
L1 + Odd	0.17	0.83	48.91	42.11
L1 + Odd	0.33	0.67	47.48	39.81
L1 + Odd	0.50	0.50	50.73	43.26
L1 + Odd	0.67	0.33	52.81	44.32
L1 + Odd	0.83	0.17	51.17	42.41
MTL Weight Tuning for Audio and Visual Tasks			Speech Recognition	
	Pretraining Dataset		LRW	
	Evaluation Dataset		SPC	
	Classifier for $(t, 512)$ features		LSTM	
	Labels		30 words	
Method	Video weight (α)	Audio weight ($1 - \alpha$)	Accuracy ($\uparrow$)	
L1 + AoT	0.17	0.83	88.74	
L1 + AoT	0.33	0.67	89.28	
L1 + AoT	0.50	0.50	90.36	
L1 + AoT	0.67	0.33	91.94	
L1 + AoT	0.83	0.17	90.79	
L1 + Odd	0.17	0.83	89.97	
L1 + Odd	0.33	0.67	88.39	
L1 + Odd	0.50	0.50	90.78	
L1 + Odd	0.67	0.33	92.17	
L1 + Odd	0.83	0.17	91.31	

Table 5.6: Comparison of different MTL weights (α) for the loss function $\alpha L_{video} + (1 - \alpha) L_{audio}$. All results are accuracies on validation sets of chosen datasets. For both tested methods, the optimal α is 0.67. This shows that the video loss term being scaled to roughly 2x the audio loss term works best for our method.



(a) Emotion recognition F1 score on CREMA under noise



(b) Speech recognition F1 score on SPC under noise

Figure 5.6: Comparison of performance of MFCCs and our methods trained using audio-only (Odd), video-only (L1) and audio-visual (L1 + Odd) self-supervision under various levels of artificially introduced babble noise.

5.4.6 Performance in various levels of noise

In order to further rigorously validate our proposed models for robustness, we investigate the performance under various levels of noise. We create noisy versions of the CREMA and SPC datasets by adding babble noise from the NOISEX database [203], while varying the SNR from -5 dB to 20 dB in steps of 5 dB. We perform a comparison between the following methods: (i) L1 using visual self-supervision, (ii) Odd using audio-only self-supervision, and (iii) L1 + Odd for bimodal self-supervision, and (iv) MFCCs. We examine how the performance varies for the methods as the level of added noise changes in the evaluation datasets.

The results for the experiments with added noise can be seen in Fig 5.6. For both datasets, we observe that self-supervision using audiovisual combination outperforms unimodal methods. For emotion recognition on CREMA, the audio features from the clean dataset give the best performance, and there is a linear degradation of performance with the increase in noise. Visual self-supervision is more effective in almost all scenarios, which may be expected as visual features are unaffected by auditory noise and can still drive robust learning of audio features. A similar conclusion can be reached for speech recognition as well. Yet audio-visual self-supervision leads to the best audio representations across all noise labels. The results for speech recognition are also significantly more robust to noise than those for emotion recognition. We can see this from the decrease in performance with increase in noise, which is not very sharp until extremely high noise levels at -5 dB and 0 dB. We can also notice a sharper degradation for the audio-only results when compared to both the video-only and audiovisual results. This suggests that audio features obtained by visual or audio-visual self-supervision are more robust to noise compared to those obtained by audio-only self-supervision.

5.4.7 Performance with various sizes of the pretraining set and downstream training set

It is also an interesting experiment to see how model performance varies with the amount of data used for self-supervised pretraining. One of the most important advantages of a self-supervised learning approach is the ability to use an arbitrarily high amount of unlabeled data to learn a good representation. But there is still a tradeoff to be made with training time and model performance. We used a subset of the LRW dataset with a total pretraining size of 112658 samples (36.3 hours of audiovisual speech) to train our full model. For this experiment, we investigate

what happens to our model if we only use a fraction of the total available data for pretraining. We use 0.2, 0.4, 0.6 and 0.8 times the dataset size for pretraining. We compare the performance between: (i) L1 for visual self-supervision, (ii) Odd for audio-only self-supervision, and (iii) L1 + Odd for audio-visual self-supervision. As before, we evaluate on the CREMA dataset for emotion recognition and the SPC dataset for speech recognition.

The results for the experiments with different pretraining set sizes can be seen in Fig. 5.7. For emotion recognition on CREMA, there is a steady and somewhat linear degradation in performance with the reduction in data in the pretraining. The best performance drops from an F1 score of 0.572 with the full dataset to an F1 score of 0.519 for 20% of the dataset. For speech recognition on SPC, there is a slower degradation for both visual and audiovisual methods, however there is a sharper degradation for the audio method with lesser training data. With 20% of the training data, the performance for the audio-only method drops to an F1 score of 0.724, which is a massive gap from the F1 score of 0.842 for the audiovisual method in the same setting. This goes on to show that the method with combined audio and visual self-supervision offers more robustness with varied amounts of pretraining data. Another observation is that a larger amount of pretraining data helps learn better features: the gain is significant initially however starts to plateau after a point (80% of the data). Fig. 5.7 also shows that with just 20% / 80% of the LRW dataset for pretraining, the audio-visual self-supervised methods achieve similar or better performance (± 0.5 for emotion, ± 0.94 for ASR) as PASE, CPC and APC which use the full pretraining set (see Tables 5.2, 5.3, and 5.4). This is a very interesting result because it shows that our proposed self-supervised methods require lesser pretraining data than other tested self-supervised methods to achieve competitive performance.

We also perform an experiment on SPC in which we simulate a limited labeled data setting by using only 10% of the training set labels for the downstream tasks. We present the results for this experiment in Table 5.7. We can see that L1 + Odd is the best performing method in the limited data setting with an F1 score of 0.897. We also notice a much sharper degradation in the performance of other baseline methods as compared to our visual and audiovisual self-supervision based methods. Our method being the best performing method in the limited data setting is also very relevant in the context of self-supervision, where the goal is to use a large unlabeled dataset to learn useful representations which are then evaluated on a smaller labeled dataset.

SPC: Limited label setting	% of labels used	
	100%	10%
Method		
MFCC	0.912 ± 0.002	0.864 ± 0.013
PASE	0.941 ± 0.002	0.811 ± 0.025
APC	0.933 ± 0.003	0.874 ± 0.011
Odd	0.919 ± 0.002	0.839 ± 0.019
L1	0.935 ± 0.003	0.877 ± 0.012
L1 + Odd	0.944 ± 0.002	0.897 ± 0.008

Table 5.7: Comparison of speech recognition performance on SPC with full training set used and with 10% of the training set used to simulate a limited labeled data setting. Results are F1 scores.

5.4.8 Comparison with finetuned and supervised versions of encoders

All the results discussed thus far were with the frozen versions of the encoders, i.e. the encoders with their fixed weights were used as feature extractors on the evaluation datasets before training a classifier. However, these encoders can also be fine-tuned to the target dataset. We also present results for finetuning in the bottom of Tables 5.2, 5.3, and 5.4. We use the weights from the L1 + Odd model as the initialization for the encoder before performing training on the target datasets in an end-to-end manner. We observe that finetuning is able to give us better results depending on the problem setting. The best result that we get on CREMA is an F1 score of 0.592, which is significantly better ($p < 0.05$) than the F1 score of 0.573 that is attained using just the frozen encoder. For SPC, we get an F1 score of 0.953 which is also the best result seen so far in all experiments. The largest observed gain comes for the Ravdess dataset, on which we get an F1 score of 0.645, compared to the F1 score of 0.565 with the frozen version. We also see gains using when finetuning on SEWA and RECOLA for continuous affect recognition. For SEWA, the valence CCC improves from 0.373 to 0.380 and the arousal CCC improves from 0.371 to 0.383. For RECOLA, the valence CCC improves from 0.448 to 0.452 and the arousal CCC improves from 0.745 to 0.764. All improvements on SEWA and RECOLA are significant with $p < 0.05$.

To motivate the utility of adopting self-supervised pretraining, it is also important to compare our method with a supervised version of an encoder with the same architecture trained from scratch directly on the target dataset (see bottom of Tables 5.2, 5.3, and 5.4). The only difference is the weight initialization, for which we use random initialization for each layer. We observe that our proposed method

(L1 + Odd) significantly outperforms ($p < 0.01$) the fully supervised method on every task. This is an important result, highlighting that a large amount of unlabeled pretraining data can be used to improve the performance on small labeled downstream datasets.

5.5 Discussion

There are a number of interesting observations from the experiments.

The audio-only self-supervised method offers competitive performance but does not always outperform the existing audio self-supervised baselines. There is also a clear observation that the proposed visual self-supervision is superior when compared to the proposed audio-only self-supervision. This is largely due to the fact that the audio features obtained by visual self-supervision are closely related to the useful information present in lip movements and facial expressions (because they must encode this information for accurate facial reconstruction during pretraining). This property is also especially useful for emotion recognition due to the correlation between emotion and facial expression information, and for speech recognition due to the information from lip movements. The results on both discrete and continuous affect recognition offer evidence that the learned representation is good for emotion.

It is also clear that the models trained using a combination of audio and visual self-supervision are able to encode complementary information from each modality to yield the best possible representations among all tested methods in this work. These representations are also the most robust in the presence of various levels of noise in the data, and offer the best performance independently of the size of the pretraining set. This is perhaps the most useful finding of this work, with the implication being that problems using audio-only speech data can benefit from using visual supervision from available audiovisual speech datasets to enhance the target representations (i.e. multimodal pretraining can be used effectively for unimodal downstream evaluation).

Another very useful finding of the work is that fine-tuning of the pretrained audiovisual self-supervised models offers significantly better performance than fully supervised training from scratch. This highlights the applicability of self-supervised pretraining in the area of affective computing. The possibility of learning more generalizable features from large unlabeled multimodal datasets before evaluating on small labeled datasets is particularly promising.

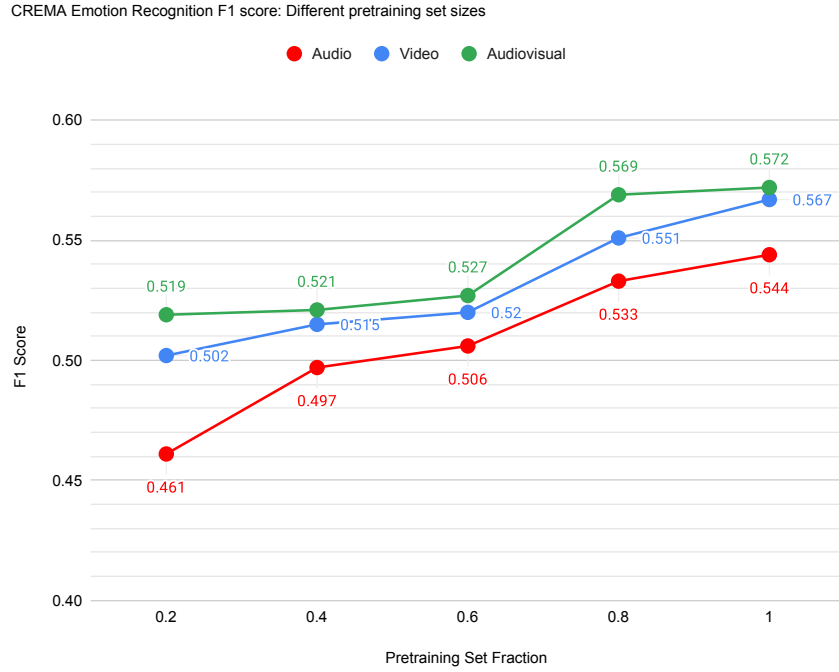
5.5.1 Limitations

A current limitation to our work is the fact that we use a nearly-frontal subset of the LRW dataset (with yaw, pitch and roll restricted to a maximum of 10 degrees each) for pretraining. This leaves out a large portion of the audiovisual dataset with profile faces which could also contain useful visual supervisory signals. There are also other larger datasets like AVSpeech [58] which could potentially yield better pretrained models. Another limitation is that we have used a very simple 2 layer BGRU with 256 hidden units as the classifier of choice for our audio classification tasks. This might not be the most optimal method or configuration. However this was chosen for simplicity. Other models such as LSTMs, Transformers, LiGRUs [166] or temporal convolutional networks (TCNs) in different configurations may yield even better results. Another possible limitation is the fact that the models in our work start from log mel spectrograms instead of raw audio (as input to the audio encoder). There is a static frequency domain transformation applied to raw audio to yield the spectrogram representation, however a more refined approach might be to use a set of trainable filters (e.g. as used in SincNet [164]) instead of static Mel filters. Another specific limitation in our experiments is that we do not account for annotation delay or do any post-processing for continuous affect recognition, both of which can potentially improve the recognition performance. In summary, the methods presented in this paper show the principle that visual and bimodal self-supervision lead to much better performances than full supervision from scratch. However, more refined approaches (that are tailored to the specific problems and datasets they are developed for) may result in even better performances than those presented here.

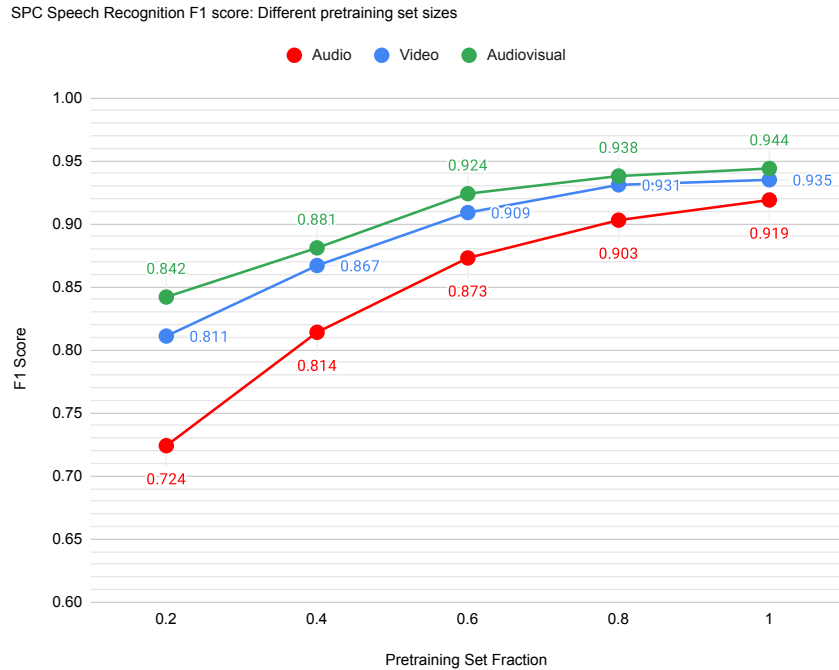
5.5.2 Future work

In this work, we have considered the interaction between the audio and visual modalities and how visual self-supervision can benefit learning of audio features. There are also other modalities that could be considered, especially the text modality. Multimodal human language is comprised of text, audio and video combined, and developing a self-supervised model that can capture the interactions between the three can be very useful. The visual pretext task that we focused on was facial reconstruction optimized by the L1 loss. This process leads to a very realistic facial animation, however this might not be the most desirable thing in order to learn the best features. Realistic reconstruction will need to capture a lot of additional information related to fine grained visual characteristics. A lot of this information

might not be useful if our end goal is simply to learn useful audio representations. Although reconstruction does give us really good performance, the question remains open to what a good alternative or additional visual pretext task might be. This work has also focused on audio features alone. It is also interesting to see how we could use audio self-supervision to guide the learning of visual speech features in an analogous way (by predicting the audio waveform from only the visual modality, like in [207]). These visual features could then be used for problems like facial affect recognition and lipreading, or even combined with our proposed audio features. There is also a promising and interesting line of work developing in audiovisual contrastive self-supervision (such as in [148, 120]), which has been demonstrated on downstream tasks like video action recognition and acoustic event detection. It would be interesting to see how such approaches perform when pretrained on audiovisual speech data and applied on emotion recognition.



(a) Emotion recognition F1 score on CREMA



(b) Speech recognition F1 score on SPC

Figure 5.7: Comparison of performance of our methods trained using audio-only (Odd), video-only (L1) and audio-visual (L1 + Odd) self-supervision for various sizes of the pretraining set (fraction of total). Total samples in LRW pretraining set = 112658. Best viewed under zoom.

Chapter 6

Joint Audiovisual Self-Supervision for Speech Representation Learning from Raw Audio

The contents of this chapter are based on the publication “Learning Speech Representations from Raw Audio by Joint Audiovisual Self-Supervision” in the workshop on Self-Supervision in Audio and Speech at the International Conference on Machine Learning (ICML), 2020 [186]. In this chapter, we show that the proposed joint audiovisual self-supervision method from the previous chapters can be applied directly on raw audio. This is performed by a slightly different network architecture (1D Resnet18 as an audio encoder instead of a GRU on Log Mel inputs). We also study downstream speech recognition performance in the presence of noise and in limited labeled data settings. We find that joint audiovisual self-supervision produces vastly superior results in situations where limited labeled data exists.

6.1 Background

The background for this chapter has been introduced in detail in Chapter 2.

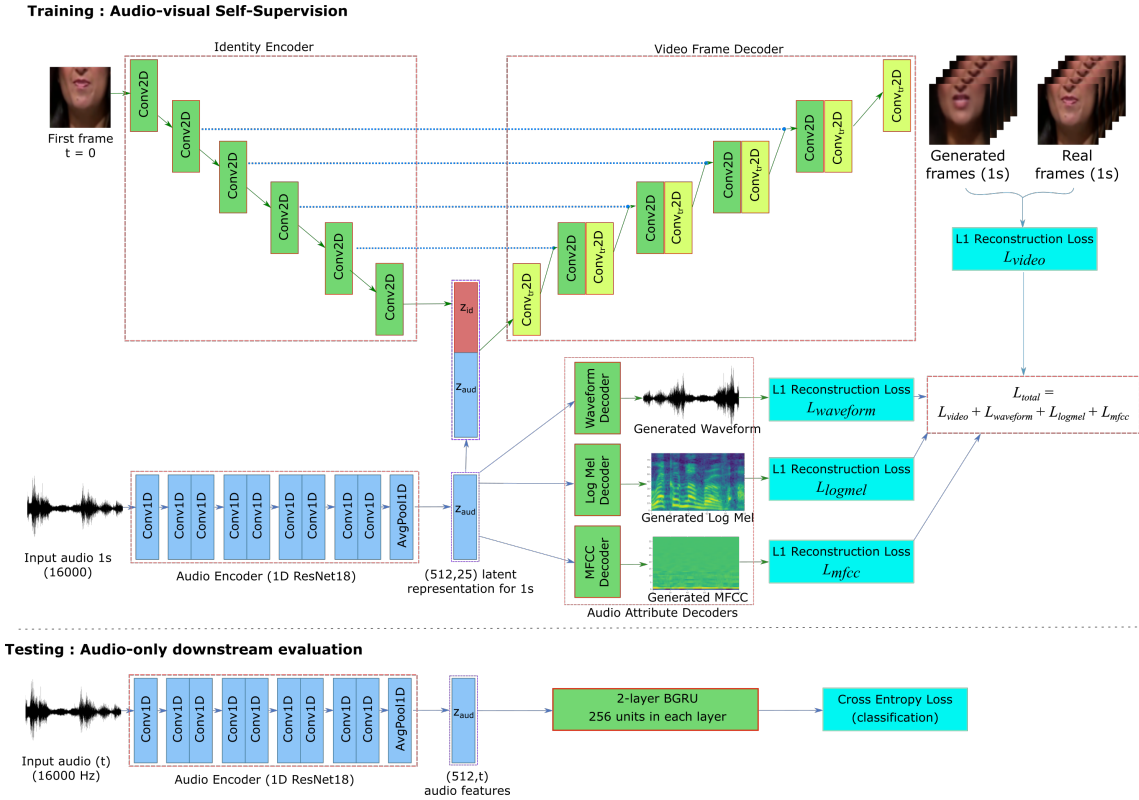


Figure 6.1: An illustration of the encoder-decoder model we use for joint audiovisual self-supervision. From an unlabeled sample of audiovisual speech, we use the raw audio waveform and the first video frame to generate a talking lip video. Lip movement reconstruction offers visual self-supervision. We also use decoders to reconstruct salient audio attributes (MFCCs, log mel, waveform) for audio-only self-supervision. By jointly optimizing the reconstruction losses for both modalities, we get joint audiovisual self-supervision. The trained audio encoder can then be used for audio-only downstream tasks.

6.2 Method

6.2.1 Audio Encoder Architecture

We use a 1D Resnet18 [80] encoder as the backbone for all of our proposed methods (detailed architecture in appendix). The encoder f_a (see Fig. 6.2 and 6.3) takes as input a 16 kHz raw audio waveform and converts it into a 512-D audio feature vector for every timestep. The output sample rate is 25 audio feature vectors per second, which matches that of 25 FPS video in the LRW dataset. This allows us to have a one-to-one mapping between the two modalities, which helps in cross-modal learning and allows us to avoid oversampling or undersampling either modality. Other contemporary self-supervised methods [8, 148] use a 2D Resnet18 audio encoder operating on mel features (operating similar to image based CNNs). However, we wanted our audio encoder to directly operate on the raw audio waveform and perform end-to-end self-supervised representation learning without starting from an intermediate feature like MFCCs or log mel spectrograms, which is why we chose a 1D Resnet18.

6.2.2 Visual Self-Supervision

For visual self-supervision, we generate a talking lip video from a still image and the corresponding audio (see Fig. 6.1 and Fig. 6.2). The model is comprised of three components: (i) the audio encoder f_a (1D Resnet18), (ii) the identity encoder f_{id} , and (iii) the frame decoder f_d . The model operates on 1 second long segments from an audiovisual speech dataset. The audio encoder f_a (Fig. 6.2 bottom-left) converts the 1 second audio sample x into a 512 dimensional embedding with 25 timesteps (z_{aud}). The identity encoder f_{id} (Fig. 6.2 top-left) is a 6 layer CNN that converts the mouth region of the first video frame x_{im} (a 64x64 image) into a 64 dimensional identity embedding (z_{id}). This embedding is replicated 25 times to match the timesteps of the audio embedding. The latent representation z is the concatenation of z_{aud} and z_{id} (as shown in Fig. 6.2). This then goes through the frame decoder f_d (see Fig. 6.2 top-right), which is a CNN that uses strided transposed convolutions to generate the video frames of the lip movements. The skip connections between the identity encoder and frame decoder help in preserving subject identity in the generated frames. An L1 reconstruction loss between frames from the generated video ($f_d(z)$) and those from the real video (y_{video}) is used to train the network. We use the L1 loss as opposed to the L2 loss to get relatively sharper reconstructions. Our model aims to pre-

Pretraining method	Self-supervision	Input	Dataset and % of Labels used			
			SPC	SPC	LRW	LRW
			100%	10%	100%	10%
MFCC	-	-	94.33	87.08	90.16	37.56
PASE [146]	Audio	Raw	95.61	83.81	93.40	1.88
APC [45]	Audio	Mel	94.87	89.91	93.97	57.41
wav2vec [175]	Audio	Raw	96.04	91.57	94.60	19.50
L1 [188]	Visual	Mel	95.11	86.43	94.45	33.43
L1 + Odd [188]	Audiovisual	Mel	95.77	90.16	94.72	67.98
Ours (A)	Audio	Raw	95.06	90.56	94.14	69.70
Ours (V)	Visual	Raw	94.38	88.31	92.18	52.99
Ours (AV)	Audiovisual	Raw	95.21	90.63	95.37	77.13
Sup. 1D Resnet18	-	Raw	93.79	81.12	90.34	13.72

Table 6.1: Results for spoken word classification (Accuracy in %) on the Speech Commands (SPC, 30 classes) [212] and the Lip Reading in the Wild (LRW, 500 classes) [41] datasets. For evaluation, a 2 layer GRU model is used on the encoder outputs for each pretraining method, before finetuning on the downstream task.

dict lip movements given only audio and speaker identity information from the first frame. In this process, the audio encoder is driven to produce useful **speech features that correlate with lip movements** (because accurate lip movement reconstruction will reduce the loss). The audio features obtained by reconstructing lip movements are likely to contain information about the speech content. Our proposed method is related to our prior work on visual self-supervision to learn audio features [187, 188, 185]. In this work, the key difference is that we use a raw audio encoder for end-to-end learning as opposed to the log mel spectrogram encoder we used in [188, 185]. Also, instead of reconstructing the full face, we focus on the mouth region which contains visual information about the speech content, which we hypothesized would lead to better representations for speech recognition.

$$z(x, x_{im}) = \text{cat}(f_a(x), f_{id}(x_{im})) \quad (6.1)$$

$$L_{video}(x, x_{im}) = |f_d(z(x, x_{im})) - y_{video}| \quad (6.2)$$

6.2.3 Audio Self-Supervision

In prior work [188], we employed temporal order based pretext task for audio-only self-supervision (predicting which of the inputs are jumbled or reversed). We wanted to examine whether it is possible to yield better speech representations

using a more refined pretext task. In this work, our methodology for audio-only self-supervision is inspired from PASE [146]. We predict three informative audio attributes: (i) MFCCs, (ii) Log mel spectrograms, and (iii) the waveform. The key difference of our method with PASE is the fact that we directly train a 1D Resnet18 encoder model on the raw audio waveform. PASE requires intermediate steps like adding speech distortions for data augmentation, SincNet filters, and a penultimate Quasi-RNN layer. We also adopt only 3 of the most informative predicted attributes from PASE for simplicity. Fig. 6.3 illustrates our method for audio-only self-supervision. The audio encoder (f_a) converts 1 second of 16 kHz input audio (x) into a 512 dimensional audio embedding (z_{aud}) with 25 timesteps (exactly the same as in the method for visual self-supervision). The audio representation is then used as input to three separate decoders (f_{mfcc} , f_{logmel} & f_{wav}) that reconstruct the desired audio attributes. We keep the decoder architectures as simple as possible in order to incentivize the important information about the audio attributes to be captured by the audio encoder. The MFCC and the log mel spectrogram decoders (Fig. 6.3 right) are both comprised of a single fully connected layer of 256 units. The waveform decoder (Fig. 6.3 top-left) is made of a transposed convolution layer followed by a convolution layer that outputs the reconstructed waveform (in an autoencoder-like fashion). We use an L1 loss between each reconstructed attribute with its ground truth (y_{attrib}) to train the model. The total loss is the sum of the MFCC loss, the log mel loss, and the waveform loss. For $attrib \in \{mfcc, logmel, wav\}$, the loss is:

$$L_{audio}(x) = \sum_{attrib} |f_{attrib}(f_a(x)) - y_{attrib}| \quad (6.3)$$

6.2.4 Audiovisual Self-Supervision

For joint audiovisual self-supervision (see Fig. 6.1), we simply combine the two proposed methods for visual-only and audio-only self-supervision. Since the same audio encoder architecture has been used in both models, we can simply use the shared audio representation as input to each of the four decoders (frame decoder, MFCC decoder, log mel decoder, waveform decoder). The total loss is the sum of the audio-only and the visual-only losses. The audio encoder (f_a) is thus trained end-to-end and is driven to produce features that contain information about each of the predicted attributes from both the audio and the visual modalities.

$$L_{total}(x, x_{im}) = L_{video}(x, x_{im}) + L_{audio}(x) \quad (6.4)$$

6.3 Experiments

Datasets The LRW dataset [41] is a large, in-the-wild dataset of 500 different isolated words primarily from BBC recordings. It is an audiovisual speech dataset and is thus appropriate for training our methods. We use a subset of LRW that has only nearly frontal videos (with yaw, pitch and roll restricted to a maximum of 10 degrees), in order to have a cleaner supervisory signal from the visual modality. This filtering leaves us with a total of around 40 hours of usable data. We use this subset of the LRW dataset for self-supervised pretraining of our proposed methods. We also use it as a spoken word classification evaluation dataset. The SPC (Speech Commands v0.01) dataset [212] contains 64,727 total utterances of 30 different words by 1,881 speakers. We use SPC also as a spoken word classification evaluation dataset.

Baselines We compare our methods against other self-supervised methods for learning speech representations. For all the baselines, we use the code (and pre-trained models) provided by the authors. We compare against PASE [146], APC [45] and wav2vec [175]. We also compare against our prior related work. L1 [188] is similar to our proposed method for visual-only self-supervision but is based on log mel spectrograms as opposed to raw audio. L1 + Odd [188] is an audio-visual self-supervised method. We use a more refined audio self-supervision approach in this work. We also compare our methods against two supervised learning baselines for audio. We use 39 dimensional MFCCs (13 coefficients, 13 deltas, and 13 delta-deltas) as the first supervised baseline. The second baseline is a fully supervised 1D Resnet18 model (same architecture as our pretrained encoders but trained from scratch directly on the evaluation datasets).

Experimental setup We evaluate all methods on isolated word classification on the Speech Commands (SPC) [212] and Lip Reading in the Wild (LRW) [41] datasets. We use a 2 layer BiGRU (with 256 units in each layer) on the encoder outputs followed by a linear layer with as many units as the number of target classes (30 for SPC, 500 for LRW). This acts as the downstream classifier and remains the same for every method. For downstream classification, we finetune the models (as shown in bottom of Fig. 6.1) for 50 epochs. The learning rate is 0.0001 for the first 40 epochs and 0.00001 for the last 10 epochs. We use the standard softmax + cross entropy loss for training. We opted to use a BiGRU for simplicity, however this can be replaced by any model that can classify variable length

Dataset	Model	Noise level (SNR)						
		-5 dB	0 dB	5dB	10 dB	15 dB	20 dB	Clean
SPC	MFCC	76.31	84.97	90.56	91.98	93.05	94.19	94.33
	Ours (A)	79.35	88.42	92.34	93.41	94.63	95.04	95.06
	Ours (V)	77.92	86.92	91.01	92.80	93.47	93.88	94.38
	Ours (AV)	79.79	88.69	92.21	93.57	94.65	95.02	95.21
LRW	MFCC	50.18	70.75	81.08	85.74	88.41	90.11	90.16
	Ours (A)	58.84	79.13	89.14	91.72	92.87	93.84	94.14
	Ours (V)	51.40	73.47	84.61	88.11	90.98	91.58	92.18
	Ours (AV)	64.63	82.59	90.08	92.09	92.91	93.87	95.37

Table 6.2: Results for spoken word classification (Accuracy in %) under various levels of introduced noise (SNR in dB). Babble noise from the NOISEX database is used to perturb the audio samples in the LRW and SPC datasets.

sequences into discrete categories (such as LSTMs, TCNs, LiGRUs [166]). The results can be seen in Table 6.1.

Results with all labels With 100% of the training dataset used, all self-supervised methods achieve comparable performance and outperform fully supervised training. On the SPC dataset, the best overall performance is attained by wav2vec with an accuracy of 96.04%, followed by our prior work at 95.77%, PASE at 95.61% and our proposed method at 95.21%. On LRW, the best performance is by our method with an accuracy of 95.37%.

Learning with fewer labels The concept of self-supervision is especially relevant to situations where labeled data is scarce. To compare the methods in such situations, we perform the same word classification experiments on the SPC and LRW datasets but with only 10% of the samples being used in the training set (the validation and test sets remain unchanged). Note that we completely omit the remaining 90% of the training set (see Tables 6.6, 6.7, 6.8 for exact split details). This leaves us with around 170 training examples per class for the SPC dataset (30 classes) and only around 20 training examples per class for the LRW dataset (500 classes). This makes the problem significantly more challenging. On SPC, there is a slight degradation in the performance of all methods. Our method attains an accuracy of 90.63% which is second to only wav2vec at an accuracy of 91.57%. On LRW, all other methods get severely affected and overfit to the small training set. Our method is the least affected and significantly outperforms all other methods with a best performance of 77.13%.

Noisy situations We also compare the performance of the variations of our method under various levels of artificially induced noise. We introduce babble noise from the NOISEX [203] database to create noisy versions of the SPC and LRW datasets. We use six levels of noise, in the range of -5 dB SNR to 20 dB SNR in increments of 5 dB. The results for the noisy datasets can be seen in Table 6.2. All our methods outperform MFCCs at all noise levels on both datasets. The joint audiovisual method is the best.

6.4 Discussion

There are multiple interesting observations from our obtained results. Audio-only supervision yields better results than visual-only supervision. However, the model trained with joint audiovisual self-supervision performs better than the models trained with unimodal audio-only and visual-only self-supervision in almost all scenarios, including noisy datasets. This highlights the utility of the complementary information encoded by visual self-supervision and demonstrates the potential of multimodal self-supervision as a useful tool in speech representation learning. Also notably, despite all tested methods being very similar in performance on the full datasets, there is a clear gap when using a small training set and our method is the best at learning with fewer labels, which is very relevant to low resource domains. This can have significant impact in problems like low resource language ASR, emotion recognition and cross-cultural ASR. Our method also significantly outperforms fully supervised training from scratch, which further motivates the utility of self-supervised pretraining for speech.

Future work This is a work in progress and there are many other speech related applications that we can evaluate our model on. In this work, we only focused on the classification of isolated words. We will also test the model on continuous CTC based speech recognition on datasets like Librispeech and TIMIT, and other tasks like speaker identification and speech emotion recognition. An especially relevant application would be low resource language ASR. There are also interesting directions to explore to improve our method. In this work, we exhibit how joint audiovisual information can be used for audio representation learning. In a similar manner, we could also utilize this cross-modal information for visual representation learning (e.g. predicting speech attributes from the visual modality). Another interesting line of work is multimodal contrastive self-supervised learning

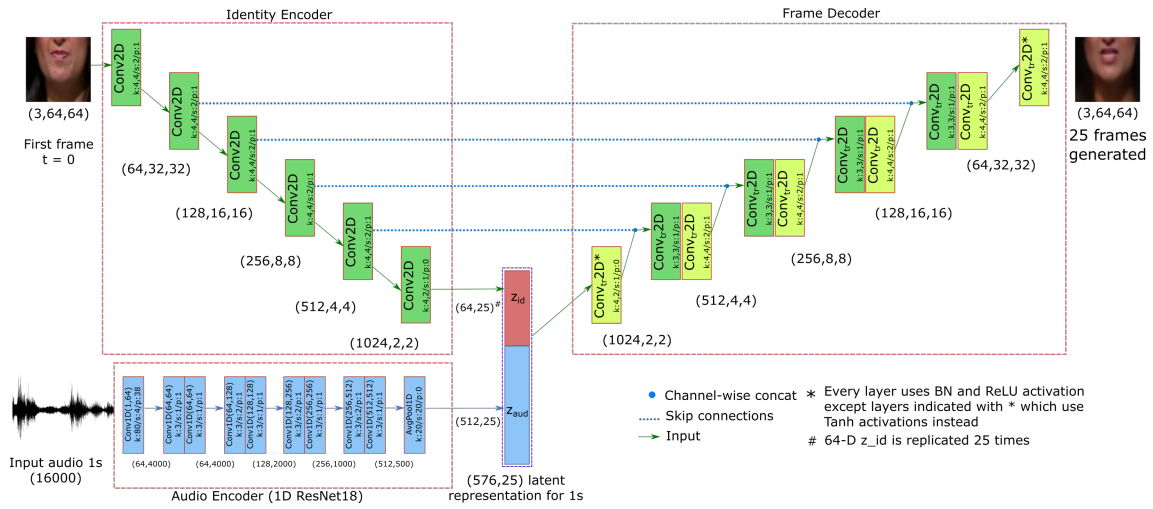


Figure 6.2: A detailed illustration of the encoder-decoder model we use for lip video reconstruction. From an unlabeled sample of audiovisual speech, we use the audio and the first frame of the video ($t = 0$) to generate a video with t frames. The model contains: (1) an identity encoder which produces a 64-D identity embedding; (2) an audio encoder which converts the input audio (t frames of 80 dimensional log mel spectrograms) into a 512-D audio embedding; (3) a frame decoder which generates video from the concatenated latent representation using transposed convolutions.

which has been demonstrated for generic audiovisual data but not for audiovisual speech.

6.5 Audio Encoders

Method	Encoder type	Parameters
PASE	SincNet + CNN + FC	5,818,020
APC	Log mel + GRU	4,105,296
wav2vec	CNN	32,537,088
L1 + Odd	Log mel + GRU	4,065,282
Ours	1D Resnet18	3,848,576

Table 6.3: Encoder type and number of trainable parameters in each of the compared methods. The architecture we use has by far the fewest parameters with 3.8M

Pretraining datasets for baselines The results in Table 6.1 for all the baseline methods (PASE, APC, wav2vec) have been computed using the public code and pretrained models provided by the authors. These baseline methods (and our

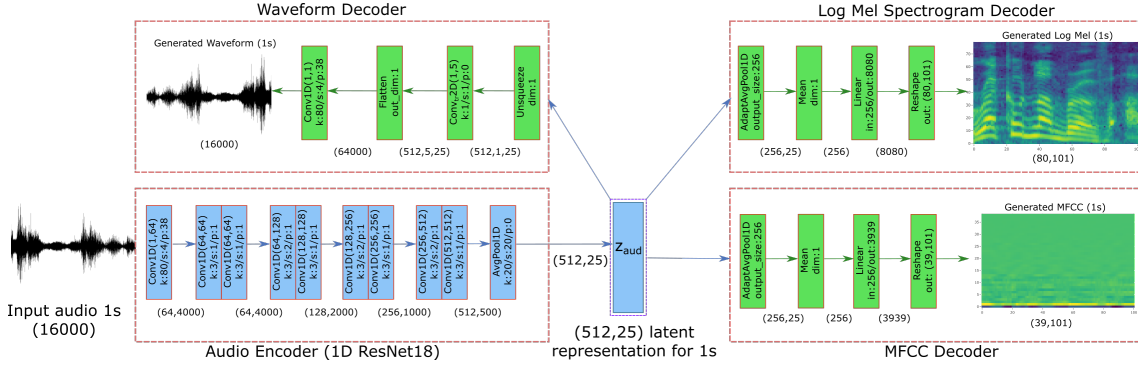


Figure 6.3: A detailed illustration of the encoder-decoder model we use for audio-only self-supervised representation learning. From an input waveform of 1 second, we predict three informative attributes: MFCC, log mel spectrogram and the waveform. The decoders are kept as simple as possible to incentivize the audio representations to capture the necessary information about the attributes.

Method	Dim.	Hz
MFCC	39	101
PASE	100	100
APC	512	101
wav2vec	512	98
L1	512	101
L1 + Odd	512	101
Ours	512	25

Table 6.4: Feature dimensionality and sample rate of each of the compared methods.

method) have been pretrained on varying amounts and types of data. For a completely fair comparison, all methods need to be pretrained with the same data. We experimented with pretraining all baseline methods on the same 36 hour LRW frontal subset that we use for our method. The results obtained with the baseline methods using this approach were either equivalent or worse to those with the public pretrained models. This shows that our model may be able to learn better representations on the same amount of pretraining data. However for the results, we use the public pretrained models which may assist with reproducibility.

Method	Pretraining Dataset	Duration
PASE	Librispeech subset	10 hours
APC	Librispeech train-clean-360	360 hours
wav2vec	Full Librispeech + WSJ	1000 hours
L1	LRW frontal subset	36 hours
L1 + Odd	LRW frontal subset	36 hours
Ours	LRW frontal subset	36 hours

Table 6.5: Pretraining dataset and duration for each method

6.6 Dataset and split details

Dataset - % labels	Split size (samples)		
	Train	Val	Test
SPC-100%	51088	6798	6835
SPC-10%	5097	6798	6835
LRW-100%	112812	5878	5987
LRW-10%	11054	5878	5987

Table 6.6: The number of data samples in each split of each dataset.

Dataset - % labels	Split duration (hours)		
	Train	Val	Test
SPC-100%	14.19	1.89	1.90
SPC-10%	1.41	1.89	1.90
LRW-100%	36.35	1.89	1.92
LRW-10%	3.56	1.89	1.92

Table 6.7: The duration (in hours) of each split of each dataset.

Dataset	Classes	n/class	t/class
SPC-100%	30	1703	28.38
SPC-10%	30	170	2.82
LRW-100%	500	225	4.36
LRW-10%	500	22	0.42

Table 6.8: The average number of samples (rounded to nearest integer) and duration in minutes of each class in the training set.

Chapter 7

Audiovisual Self-Supervised Learning of General Sound Representations

So far in this thesis, we have studied how visually guided self-supervision can be used to drive the learning of better audio representations in the context of audiovisual speech data. However, audiovisual data as a whole is much more diverse than just speech. Even though speech forms a critical subset of this, it fundamentally differs from other general audiovisual data in that it is highly structured in both the audio (composed of phonemes of a certain specific duration that have a limited vocabulary for a given language) and the video (audiovisual speech data is exclusively talking heads, with very ‘clean’ lip movements compared to an exponentially wider gamut of information in other general audiovisual data) modalities. Specific methods that may work for representation learning of speech data (e.g. the proposed speech-driven facial reconstruction method) may not work for general audiovisual data (e.g. audio event classes from YouTube such as in the Audioset [69] dataset).

In this chapter, we investigate and propose a different body of methods for audiovisual self-supervised learning that can be more generally applied to learning sound representations (as opposed to only speech representations). We first examine very recent state-of-the-art methods based on momentum-based contrastive learning (MoCo [81], GDT[148]), clustering-based pseudolabel prediction (XDC [8]), simple siamese contrastive learning (SimSiam [36]) and dimension-wise cross-correlation maximization and independence (Barlow Twins [222]). We then investigate how improvements and modifications can be made to these methods to

make them suitable for audio and audiovisual sound event representation learning. We propose a total of 4 new/adapted methods:

1. **AVCC: Audiovisual Contrastive + Clustering:** This method combines momentum based cross-modal learning from GDT [148] and cross-modal pseudolabel based clustering from XDC [8]. It introduces additional positive pairs into the audio and visual memory banks by using clustering that help learn better sound representations.
2. **A-V SimSiam:** This method is a straightforward extension of the SimSiam [36] method for natural images to the audio-only and audiovisual settings.
3. **Audio-only Barlow Twins:** This method is a straightforward adaptation of the Barlow Twins [222] method for natural images to the audio modality.
4. **Audiovisual sound source localization based Barlow Twins:** This method proposes audiovisual sound-source localization as a pretext-task to guide the learning of better audio features that are better correlated with the spatiotemporal visual features from the sound producing regions of the video modality. To the best of the author’s knowledge, this is the first method to propose using self-supervised audiovisual sound source detection for representation learning (just like the previous parts of the thesis described using lip regions in video to learn better speech representations). A combination of audiovisual synchronization and the Barlow Twins loss are used for training the model which produces the best overall representations out of the tested methods.

We evaluate the proposed representations on three different tasks: (a) environmental sound classification on ESC50 [156], (b) speaker identification on Voxceleb1 [135], and (c) audio event detection on the balanced set of Audioset [69]. We present detailed and exhaustive ablation studies to determine the best variation of each method. Our proposed methods outperform the existing methods on these tested datasets.

7.1 Background

While Chapter 2 introduces detailed background in self-supervised learning for images and audiovisual data, we perform a more detailed introduction of some existing methods in this chapter. These existing methods form the basis of many of our proposed methods for audiovisual self-supervised learning for sounds.

7.1.1 Related work in Visual and Audiovisual Self-Supervision

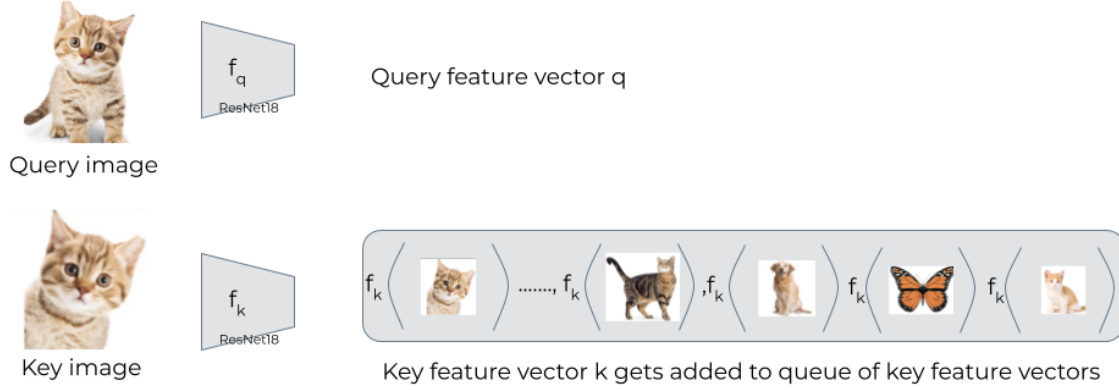
A seminal recent work in self-supervised learning is MoCo [81, 82] which stands for Momentum Contrast. While contrastive learning methods have been gaining popularity recently, a crucial bottleneck was the ability to fit a limited number of samples in a batch on the training hardware (typically GPUs). The idea behind contrastive learning is fundamentally to reduce the representation distances between the positive elements in a batch and to increase the representation distances between negative pairs in the batch. This can also be interpreted as a classification problem with the standard cross-entropy loss (the MoCo code does exactly this), in which the problem essentially becomes ‘classifying which of the candidate elements is a positive example’. The issue however became being able to fit only a limited number of negative examples per batch on the GPU. Intuitively, this leads to potentially worse representations because picking the correct positive sample from a smaller number of candidate samples (e.g. 1 out of 8) is a much simpler problem than picking the correct one from a much larger batch (e.g. 1 out of 100000). Solving the harder problem here should lead to better representations because the network is incentivized to find better discriminative features for the chosen data points. MoCo addresses this issue by introducing the idea of using a ‘memory bank’ or a feature vector queue instead of just computing the forward pass for every sample in the candidate batch. Because only the computed features (typically of dimensionalities like 512 or 1024) need to be stored instead of doing the whole forward pass again, this allowed the scaling up of the number of negative samples that could be used to compute the contrastive training loss. Figure 7.1 shows the idea of MoCo in depth.

GDT [148] extends the idea of MoCo (which was primarily demonstrated on ImageNet) to the audiovisual domain. The fundamental idea for representation learning remained the same (using memory banks and paired query and key encoders), however the process is done in a cross-modal fashion. The audio query encoder is trained by learning to predict the correct corresponding video key, and the video query encoder is trained by learning to predict the correct corresponding audio key. This leads to better representations being learned for both modalities. Figure 7.2 illustrates GDT and how it maintains a total of 4 networks (the audio and video query and key encoders) while maintaining two separate memory banks (one each for audio and video).

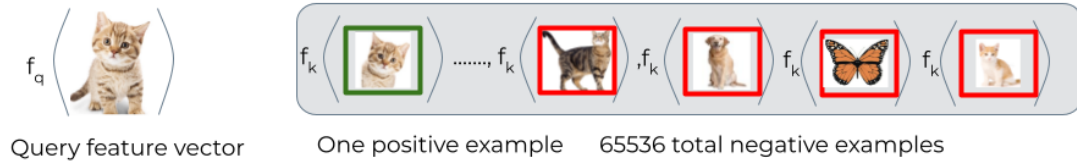
Both MoCo and GDT are invariance-based methods which operate on separate augmented samples (with standard augmentations being used for both video and

MoCo

1. Encoding different augmentations of the same image



2. Picking the correct example from the queue of feature vectors



For 512-D query feature vector q , compute dot product with all elements in the queue

Aim is to make this inner product highest for the positive example and low for negatives

Using contrastive loss, the representations of the positive pair come closer together and those for all negative pairs are pushed further away

3. Weight update

Loss = Cross entropy loss of classifying the positive example correctly



Figure 7.1: A detailed explanation of MoCo (Momentum Contrast) method by He et al. [81, 82]. The memory banks are useful in increasing the difficulty of the instance discrimination task by scaling the number of negative examples, which leads to better representations being learned when correctly classifying the positive pairs.

GDT -Extending MoCo to audiovisual setting

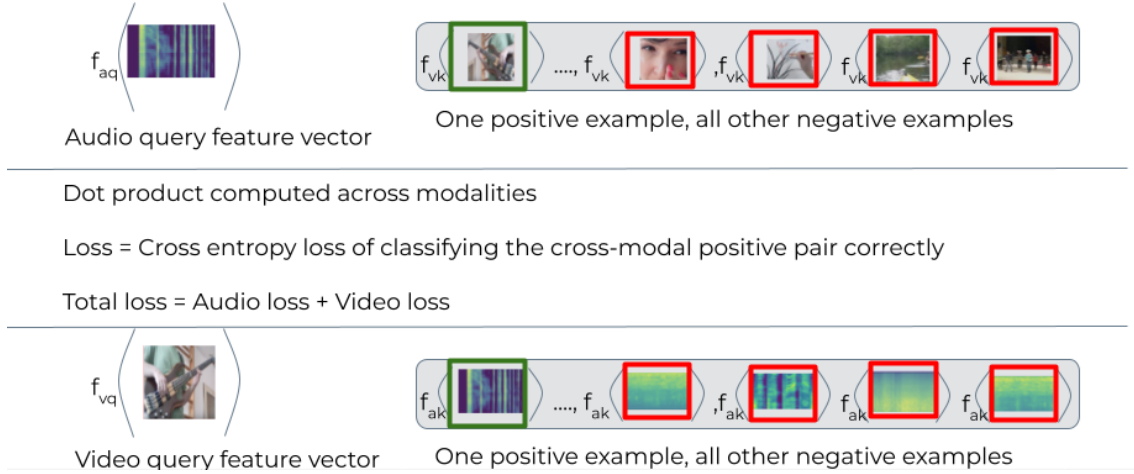


Figure 7.2: An illustration of the work GDT by Patrick et al. [148]. The method can be conceptually thought of as an extension of MoCo [81] to an audiovisual setting where the memory banks and the contrastive learning process is performed in a cross-modal manner.

audio) of each data point.

XDC (Cross-modal Deep Clustering) [8] is another recent method that extends the idea of Deep Clustering [26] to the audiovisual domain. The encoders of each modality are trained using a cross-entropy loss to classify the cluster IDs (pseudolabels) of the representations produced by the encoder of other modality. Before every training epoch, the features produced by both encoders are re-assigned cluster IDs by using k-means clustering. As training progresses, the encoders iteratively produce better features that can successfully do the cluster prediction task by using the correspondences across the two modalities. Figure 7.3 illustrates XDC.

Our first proposed method in this chapter (**AVCC: AV Contrastive + Clustering**) conceptually builds on GDT and XDC and is introduced in detail later (in Section 7.2.1).

The next method for self-supervision on natural images that we extend to audiovisual data is SimSiam [36], which stands for Simple Siamese representation learning. The crucial advancement that SimSiam made was show the usefulness and necessity of the stop-gradient operation (which under suitable parameters prevents representation collapse). The network is trained to bring the representations of two separate augmentations of the same image closer to each other, with one branch (also containing a critical predictor MLP head on top) being trained by gradient descent (or another optimizer), while the other is not. This can be seen

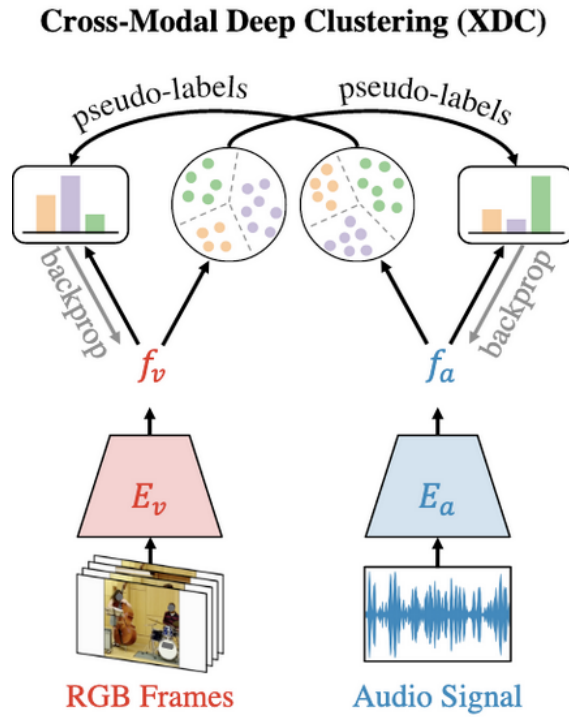


Figure 7.3: An illustration of the work on Cross-Modal Deep Clustering (XDC) by Alwassel et al. [8]. The representations for each modality are trained by predicting the pseudolabels (cluster IDs) that have been assigned to the representations of the same data clip for the other modality.

as a simpler variation of other methods like MoCo [81], BYOL [74] and SWaV [28] and is effective for learning visual representations. We extend it to the audiovisual setting as described later in Section 7.2.2.

The final method that we build upon is Barlow Twins [222]. All of the previously discussed methods for contrastive learning had an issue of them being very brittle to hyperparameter changes and prone to representation collapse. Barlow Twins addressed this issue very elegantly by using a cross-correlation based loss function, in which the target was to make the cross-correlation matrix as close to identity as possible for two augmentations of the same data sample. The loss function is given by:

$$L_{BT} = \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{i \neq j} C_{ij}^2 \quad (7.1)$$

where C is the cross-correlation matrix computed using the respective output feature vectors of the two separate augmentations that are input to the encoder network and λ is the scaling factor to determine the relative importance of each term. The first term enforces invariance based self-supervised learning, i.e. tries to make the representations of the two augmentations similar (by making the diagonal elements of the cross-correlation matrix high). The second term acts as a redundancy reduction term, which tries to maximize the independence between each dimension of the output representation and disincentivizes representation collapse or trivial solutions. Thus, the cross-correlation matrix being the identity matrix is the target of the process. This elegant formulation outperforms the previously proposed methods while also being much easier to train and understand. We extend the idea of Barlow Twins to the audio-only and audiovisual settings in Section 7.2.3.

Audiovisual sound-source localization

Chapter 2 introduces work on self-supervised sound source localization in detail.

In this chapter, we leverage self-supervised audiovisual sound source localization (using an adaption of the AVObjects method by Afouras et al. [5]) to extract spatiotemporal features from the sound producing regions of the video modality to guide the learning of better audio representations. The AVObjects method works by operating on a spatial output feature map (as opposed to a typical single output feature vector) for the video modality, and trying to learn the location of the sound producing regions to be that which maximizes the dot product between

the audio and video features. The training is done by using real data as positive pairs and artificially misaligned audio and video as the negative pairs. The real pairs force the learning of the correct sound producing regions, because only the sound producing regions will have consistently high correlation with the audio embeddings, whereas noiseless and desynchronized regions will not. We adapt this method and combine it with the Barlow Twins objective, and describe it in detail in Section 7.2.3.

7.2 Self-Supervised Learning for Generic Audiovisual Data

This section describes the proposed methods for audiovisual self-supervised learning in detail. All of these methods are extensions and/or combinations of the methods discussed in the previous section.

7.2.1 AVCC - Pseudolabel based clustering with Memory Banks

Our first proposed method in this chapter is called **AVCC** (Audiovisual Contrastive + Clustering). The method conceptually builds on GDT [148] and XDC [8], while trying to simultaneously address issues with each.

Figure 7.4 provides an overview of our method. It is composed of paired audio and video key and query encoders very similar to GDT. The learning mechanism is also very similar to GDT (query encoders trained using cross-entropy loss, key encoders trained by momentum updates). However, the key difference compared to GDT is the using additional positive examples (instead of just one positive pair coming from an augmentation of the same clip). The additional positives are mined using clustering-based pseudolabels (similar to XDC). The rationale behind getting additional positive examples is that there may be elements in the large feature queue (we use a size of 65,536) that are incorrectly treated as negatives (e.g. there may be multiple videos of motorcycles in the video queue with the corresponding engine sounds in the audio queue, however only augmentations of the same exact clip will be treated as positives). Training the models to only predict the exact same sample as positive is really good to learn invariances at a sample level (and perhaps to very well represent near duplicates), but the introduction of other positives that belong to the same cluster may have the potential to learn better semantic representations.

Ours (AV Contrastive + Clustering): Positives belonging to same cluster

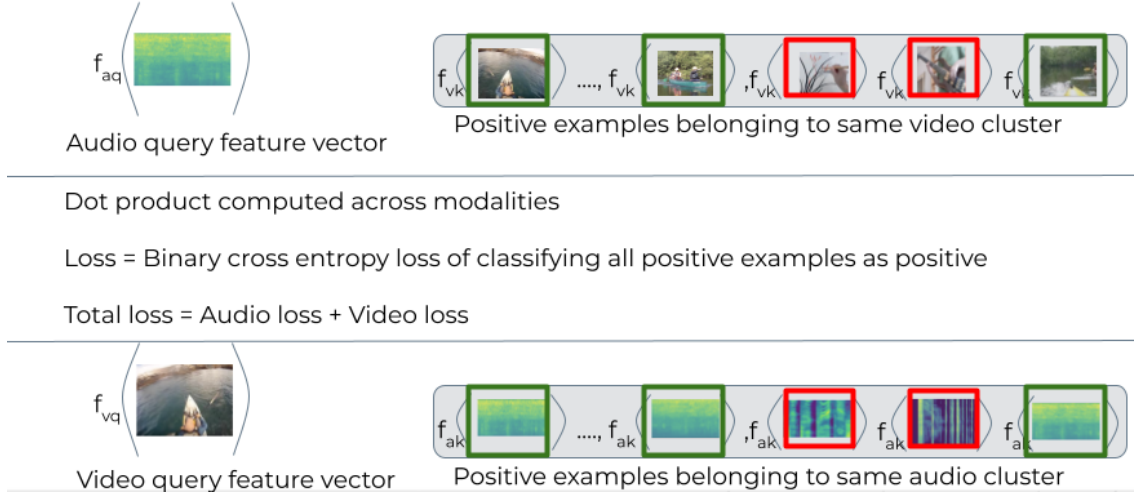


Figure 7.4: An illustration of our proposed method AVCC (AV Contrastive + Clustering) which combines the concepts of GDT [148] and XDC [8]. The intuition behind it is to introduce additional positive labeled samples into the memory banks that come from the same cross-modal clusters (in order to alleviate the bad effect of samples that are either visually or aurally similar but are treated as negative pairs if only doing GDT style training). Clustering in the representation space (similar to XDC) allows us to fetch additional positive examples that are similar to the current query sample albeit being from different data clips.

Figure 7.5 provides a detailed illustration of our method. We train a total of 4 networks: the audio query encoder f_{aq} , the video query encoder f_{vq} , the audio key encoder f_{ak} and the video key encoder f_{vk} . Before every epoch, the key encoders for each modality compute features for every sample in the training set, which we subsequently cluster using k-means clustering into one of k pseudolabels. During a single forward pass, the query encoders compute query representations (q_a and q_v) for both modalities. This is followed by computing a dot product of the respective query vectors with every key vector in the corresponding queue. This yields the audio and video logits (as shown in Fig. 7.5), and we also have their corresponding pseudolabels (cluster ID's). We then compute the loss using class-wise binary cross-entropy (Pytorch BCEWithLogitsLoss) on the multi-hot feature vector with all samples that have the same pseudolabel as the query sample being treated as positives.

For all the variations of our proposed model, the video encoder network is R(2+1)D-18 [200], and the audio network is a Resnet-18 [80] that operates on log mel spectrograms with 80 frequency bins and 100 timesteps per second. We preprocess each dataset to be 30 frames per second for video, and we use one seconds worth of time-aligned audiovisual data from a given clip with a random

AV Contrastive + Clustering

Before every epoch, perform clustering and assign each audio/video feature one of K pseudolabels

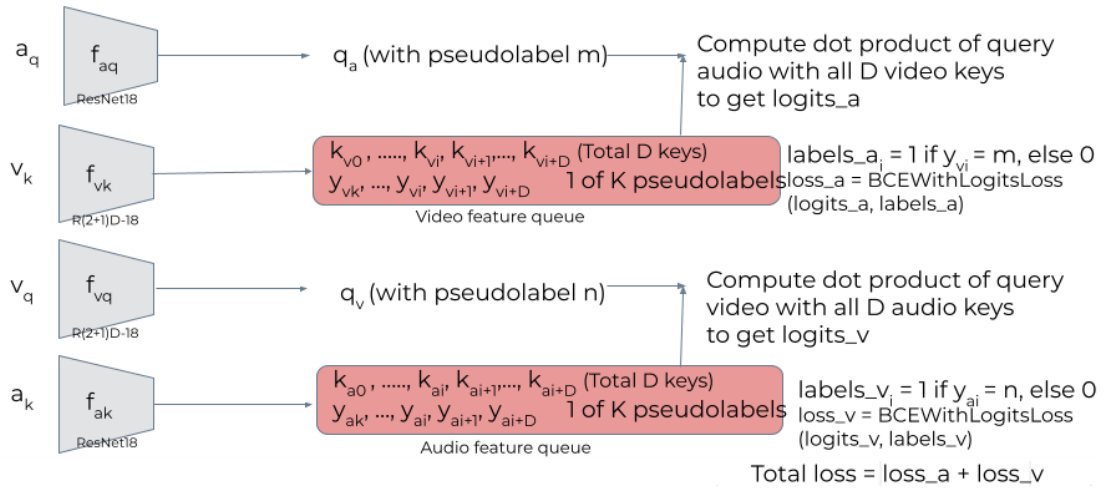


Figure 7.5: A detailed illustration of our proposed method AVCC (AV Contrastive + Clustering). There are a total of 4 networks: the audio query encoder, the video query encoder, the audio key encoder, the video key encoder. Before every epoch, k-means clustering is performed for the features of each modality to assign them each data sample one of k discrete labels. The audio query encoder is then trained to predict the positive pairs from the memory bank of video features that is populated by the features produced by the video key encoder. Vice versa, the video query encoder is trained by predicting the positive pairs from the features produced by the audio key encoder. The loss at each step is a summed class-wise binary cross-entropy loss for a multi-label classification problem, and the total loss is the sum of the audio and video query losses. The query encoders are trained by gradient descent and the key encoders are trained by the momentum update similar to MoCo [81].

start time. We use SGD to train these models with a base learning rate of 0.01 that is scaled by the number of GPUs, with a weight decay of $10e-5$ and momentum of 0.9. We use a batch size of 12 per GPU and train our models in a distributed manner on 64 total GPUs across 8 compute nodes, for a total effective batch size of 768. Note that the training parameters and augmentations are as close to the original GDT work [148] as possible to ensure the fairest possible comparison. We discuss the ablation studies (including number of clusters, pretraining dataset type and size) in Section 7.3.2 and downstream results in Section 7.3.3.

7.2.2 Audiovisual SimSiam

The second proposed method in this chapter is an audiovisual version of the SimSiam [36] method. Figure 7.6 illustrates the possible variations of the method. The A-A SimSiam and V-V SimSiam are extremely straightforward, with the only changes to the original method being the network architecture and the input data. For the A-V SimSiam variation, we use two separate parallel cross-modal SimSiam branches, with the first one having the stop-gradient operation applied on the audio branch, and the second one having it applied to the video branch. The parameters for the predictor MLP remain the same irrespective of the modality on which it operates. We set both the feature dimensionality and the predictor MLP dimensionality to be 512 for all model variations.

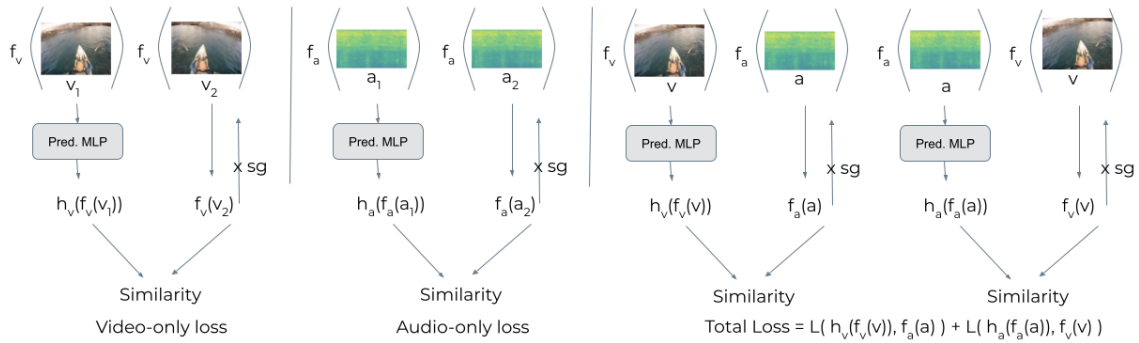


Figure 7.6: An illustration of our proposed variations of the SimSiam [36] method to learn audio, visual and audiovisual representations. The video-only (V-V SimSiam) and audio-only (A-A SimSiam) variations are straightforward extensions of the original method in which two separate augmentations of each input sample are trained to produce similar representations. The audiovisual extension (A-V SimSiam) computes the loss in a cross-modal manner and makes the representations of the two modalities from the same data point come closer to each other.

We modify the codebase released by the SimSiam authors to incorporate using

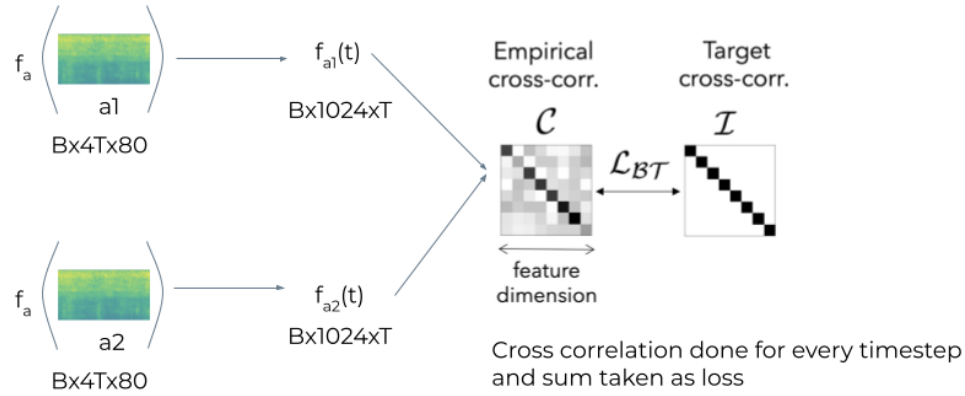
audiovisual inputs. Similar to AVCC, we use R(2+1)D-18 [200] as the video encoder network, and a Resnet-18 [80] as the audio encoder network. We train using 1 second snippets randomly sampled from 30 fps videos from the chosen pretraining datasets. We also use the exact same augmentations for both modalities as we did in AVCC. We use SGD (with the same parameters as used by the SimSiam authors) with a base learning rate of 0.05 scaled by the number of GPUs, weight decay of $10e-4$ and momentum of 0.9. We use a cosine learning rate schedule as used in the original SimSiam work. Similar to AVCC, we also use a batch size of 12 per GPU and train on 64 GPUs distributed across 8 training nodes. We train for a total of 100 epochs. We discuss results obtained by the proposed extracted features in Section 5.4.

7.2.3 Audiovisual Barlow Twins

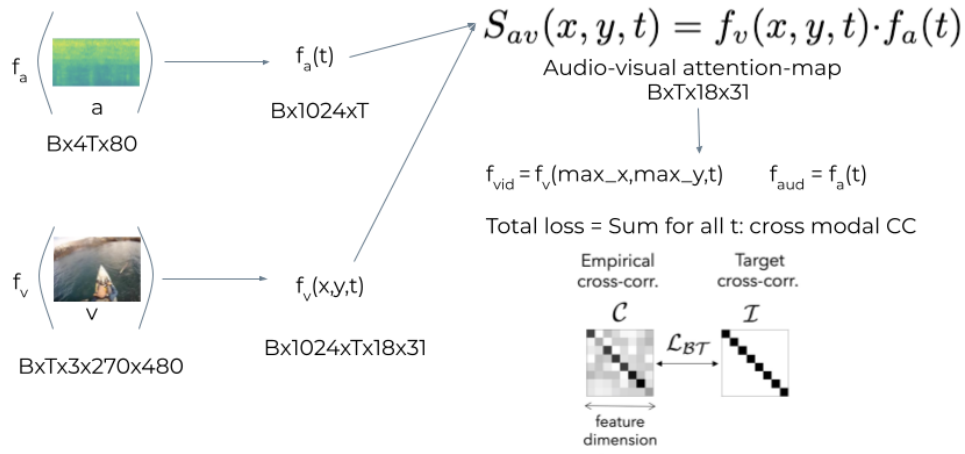
We then propose three variations of the Barlow Twins [222] method for audio and audiovisual representation learning (all of which are illustrated in Figure 7.7):

A-A Barlow Twins: The first variation is a straightforward extension of the original Barlow Twins method, with the only change being the input data being converted to augmentations of audio data samples (Audio-Audio Barlow Twins). We use a Resnet-18 [80] encoder operating on log mel spectrograms containing 80 frequency bins and 100 timesteps per second. We use the same standard audio augmentations used for the other proposed methods (SpecAugment, warping, noise addition, time-domain volume jittering) to create the two views (augmentations are based on GDT [148]). For training the Barlow twins method, we audiovisual data with 25 frames per second (similar to AVObjects [5]). For training the audio-only variation, we convert an input audio frame of $4T$ timesteps into an output audio frame of T timesteps with varying representation dimensionality (for our implementation, $T = 25$ for 1 seconds worth of data). We use a projector MLP on top of the encoder to yield the eventual audio representations. The Barlow Twins loss was introduced earlier in Equation 7.1. We compute the cross-correlation based loss between the extracted features from the separate audio augmentations, and we compute the loss for all T timesteps before summing to get the final loss used for training the model.

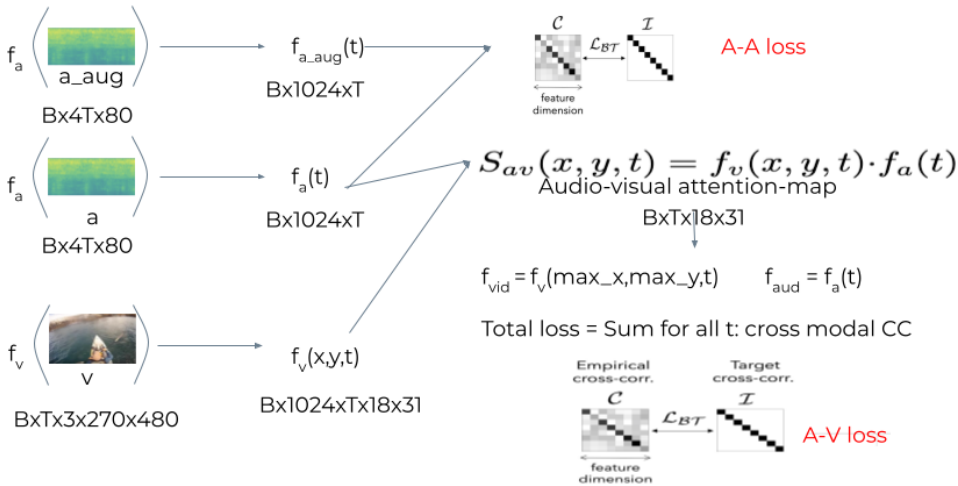
A-V Localization Barlow Twins: As described earlier, we adapt the AVObjects method by Afouras et al. [5] for sound source localization in general audiovisual data. Once we have a rough spatiotemporal attention map of the sound producing regions, we use the Barlow Twins objective to perform representation learning in



(a) A-A Barlow Twins



(b) A-V Localization Barlow Twins



(c) A-A + A-V Localization Barlow Twins

Figure 7.7: The three separate variations of our proposed methods that build on the original Barlow Twins work [222]. (a) A-A Barlow Twins operates on two separate augmentations of the same audio sample. (b) A-V Localization Barlow Twins incorporates self-supervised sound source localization (Afouras et al. [5]) to maximize cross-correlation of the input audio with the corresponding sound-producing region in the video. (c) A-A + A-V Localization Barlow Twins combines the methods

a cross-modal fashion. The audio network and feature parameters are the same as that for A-A Barlow Twins. For the video modality, we use resize and resample all videos in the respective datasets to 480x270 resolution, and we use a modified VGG [189] network similar to AVObjects [5] that converts input video of dimensions $[B, T, 3, 270, 480]$ into an output spatiotemporal video feature of dimensions $[B, D, T, 18, 31]$, where B is the batch size, T is the number of timesteps (25 timesteps for 1 second for a 25 fps video), D is the chosen output dimensionality. The 18x31 output spatial resolution (i.e. 558 total output grid points) is consistent with AVObjects. We first train just the AV source localization module using artificially misaligned audio and video as negative pairs and the true audio and video as positive pairs similar to AVObjects. We then introduce Barlow Twins style training with the added projector on top of the audio and only the maximally attended spatial video location. This helps the audio model get guided by visual supervision from the sound producing regions of the video modality, to hopefully drive the learning of better audio representations that can better capture the cross-modal correspondence. To the best of the author’s knowledge, this is the first time self-supervised AV sound source localization has been proposed for audio representation learning. This is also very conceptually similar to the idea from the earlier chapters which used speech-driven facial animation to focus on the lip regions to guide the learning of better audio speech representations.

A-A + A-V Localization Barlow Twins: Finally, we combine the above two proposed methods and train them jointly using an equally weighted multi-task loss. The intuition is that jointly training the two methods will provide advantage of the audio-only invariances being learned by the A-A branch, and the spatiotemporally sound producing regions from the video providing better cross-modal correspondences and complementary information in the A-V Loc branch.

Training settings and parameters: For all three proposed methods that are variations of Barlow Twins, we use the same training hyperparameters. We use SGD with an initial learning rate of 0.01 which is scaled by the number of GPUs. The weight decay is set to 10e-5 and the momentum is 0.9. We use a Cosine learning rate schedule. The parameter λ in the Barlow Twins loss (Eq. 7.1) is set to 0.005 following the original Barlow Twins paper. We perform distributed training on 64 GPUs across 8 compute nodes. As with the other methods, we use 1 seconds worth of data at a time as input to the models, and we use a batch size of 12 per GPU, for an effective batch size of 768. Ablation studies are presented in Section 7.3.2 and downstream audio results are presented in Section 7.3.3.

7.3 Experiments and Results

This section introduces the experimental setup (including datasets for pretraining and evaluation, encoder and audio feature design choices, ablation studies) and the associated results. We present results on a number of downstream audio recognition tasks.

7.3.1 Datasets

We use a number of different datasets for large-scale pretraining and for downstream evaluation.

Pretraining datasets

For large-scale audiovisual pretraining, we use the Kinetics-400 [95], Kinetics-700 [190], and the Audioset [69] datasets.

The Kinetics-400 dataset is the smallest of these, and has $\sim 240k$ total clips of 400 different classes of human actions in videos.

The Kinetics-700 dataset is a larger dataset with more human action classes. The total size of this is $\sim 650k$ total clips out of which we randomly select 480k total clips (so as to be left with a dataset double the size of the Kinetics-400 dataset while having a more diverse set of action classes). We also ensure to not include any clips that do not have an audio stream in the chosen subset.

The Audioset dataset is a significantly larger dataset with $\sim 2M$ total clips. The crucial benefit of Audioset for our particular task (audiovisual representation learning) is that it was specifically curated to have rich audio information (whereas the other two datasets are more focused on rich visual information). This intuitively makes Audioset a more appropriate candidate for a pretraining dataset than the other two. However, the full training set of Audioset is prohibitively expensive to store and process. Therefore, we used a subset of $\sim 240k$ videos (the exact same size as Kinetics-400) similar to Patrick et al. [148]. This also enables us to analyze the relative benefit of using an audio specific dataset to perform large-scale audiovisual self-supervised training vs using a more visually specific dataset.

Video clips in all 3 of these datasets are exactly 10 seconds long.

Note that the originally curated and labeled datasets keep shrinking over time due to YouTube videos being unavailable for a variety of reasons (being taken down by the uploaders, made private, copyright strikes, and others). Thus it is

prudent to use a subset of the total available dataset that is an appropriate size for the analysis that we need to perform.

Downstream evaluation datasets

For downstream evaluation of our pretrained models, we use the ESC50 [156] dataset for environmental sound classification, the Voxceleb1 [135] dataset for speaker identification, and the balanced set of Audioset [69] for audio-only and audiovisual sound event detection. These are all standard evaluation datasets for the chosen problems.

ESC50 [156] is a collection of exactly 2000 audio clips from 50 discrete environmental sound classes. Each class has exactly 40 samples. We use the same split for evaluation as used by other works like Patrick et al. [148] and Alwassel et al. [8] with 1200 training, 400 validation, and 400 test samples.

The VoxCeleb1 [135] dataset contains audiovisual data from 1251 different celebrity speakers. We perform audio-only downstream speaker identification. There are $\sim 140k$ train, $\sim 8k$ validation, and $\sim 8k$ test samples in the dataset.

For Audioset, using the balanced set enables us to use a downstream training set that has no overlap with the original training set which we use for self-supervised pretraining. There are $\sim 22k$ train, $\sim 24k$ val (the original validation set of the full Audioset), and $\sim 20k$ test (the original test set of the full Audioset) samples in the dataset.

7.3.2 Ablation Studies

To better understand the implications of certain parameters for the proposed methods, we perform the following ablation studies:

1. Ablation study for optimal value of number of clusters (k) for k-means clustering for the AV Contrastive + Clustering method
2. Ablation study for the size of pretraining dataset and type of pretraining (comparison with scratch and supervised) for the AV Contrastive + Clustering method
3. Ablation study for the dimensionality of features used to compute the cross-correlation loss for the Barlow Twins method

Ablation Study 1: AV Contrastive + Clustering (AVCC)

Evaluation Dataset		Audioset Balanced					
Labels		527 sound classes					
Metric		mAP ($\uparrow$)					
		Number of clusters k					
Method	Pretraining	64	128	256	512	1024	
AVCC - Audio only	Kinetics 400	0.215	0.220	0.221	0.219	0.218	
AVCC - Audio only	Kinetics 700	0.218	0.222	0.224	0.224	0.223	
AVCC - Audio only	Audioset 240k	0.230	0.238	0.242	0.243	0.241	
AVCC - Audiovisual	Kinetics 400	0.252	0.252	0.256	0.255	0.255	
AVCC - Audiovisual	Kinetics 700	0.251	0.256	0.254	0.255	0.257	
AVCC - Audiovisual	Audioset 240k	0.263	0.270	0.269	0.273	0.270	

Table 7.1: Ablation study for value of k in AVCC, with results presented on the balanced set of Audioset. The results denoted as audio-only are for unimodal audio recognition, whereas those denoted audiovisual are fused multimodal evaluation by concatenating the audio and visual embeddings. We find that a number of clusters between 256 and 1024 is the optimal choice, which varies depending on the pretraining dataset and for the evaluation setting.

AV Contrastive + Clustering: Value of k for k -means clustering

A critical parameter in the AVCC method is the number of clusters k used for k -means clustering. Prior work like DeepCluster [26] and XDC [8] have studied the same and have found that varying this according to the chosen dataset can lead to different results depending on the inherent number of concepts that exist in the chosen dataset. We vary the number of clusters from the set [64, 128, 256, 512, 1024], which is the exact same setup used in the construction of the ablation study of XDC. These numbers also make sense when being compared on the chosen pretraining datasets: Kinetics-400 and Kinetics-700 have 400 and 700 classes respectively, and the chosen cluster numbers in the study can both oversegment and undersegment the available classes. The same applies for the Audioset-240k set, which has 527 total classes.

We present results for this ablation study in Table 7.1. When pretraining on Kinetics-400, we find that the optimal number of clusters is 256 (with an audio-only mAP of 0.221 and audiovisual mAP of 0.256 being the best results). The performance for both higher and lower number of clusters reduces for this setting. When using Kinetics-700 (double the size of Kinetics-400), for audio only evaluation we find very little difference between any of 128, 256, 512, or 1024 clusters; with 256 and 512 being very marginally the jointly best performing with

an mAP of 0.224. For the audiovisual Kinetics-700 variant however, 1024 clusters is the best choice with an mAP of 0.257. This is interesting because it shows that a multimodal setting may require a higher number of clusters to represent more positives accurately in the memory banks. For Audioset-240k, we find that 512 is the best number of clusters for both audio and audiovisual evaluation, with respective mAPs of 0.243 and 0.273.

k seems to matter in a way that it should not be “too low”. For the higher values of k , sometimes the dropoff from the optimal values of k is not that high. However for lower values of k (like 64), there is a sharper dropoff in performance. The choice of k will largely map to how much we will “oversegment” or “undersegment” the implicit label distribution of the data in a self-supervised way. Choosing a value that is closer to the number of classes in the pretraining dataset seems to give better performance than choosing a lower number.

Based on the results of this ablation study, for all subsequent experiments we pick the best performing variants of the model. For Kinetics-400: we pick the model trained with 256 clusters. For Kinetics-700, we pick the model trained with 512 clusters for audio-only evaluation and the model trained with 1024 clusters for audiovisual evaluation. For Audioset-240k, we pick the model trained with 512 clusters.

AV Contrastive + Clustering: Size and type of pretraining dataset

Now that we know the best number of clusters to use for each AVCC model variation, we compare the performance of these for the size of pretraining set and the type of pretraining done with the same model backbones for audio-only downstream recognition. The audio model backbone used for all of these is a 2D Resnet-18 operating on the same type of log mel spectrograms with 80 frequency bins and 100 timesteps per second. We compare results obtained by each of the models on the different types of pretraining datasets: the 240k sized Kinetics-400, the 480k-sized Kinetics-700, and the more audio-specific 240k-sized Audioset. We also compare the performance with training the same encoder architecture from scratch, and the same encoder architecture trained in a supervised manner on each of the 3 chosen pretraining datasets, and also one variation with the encoder trained on ImageNet. We present results for this ablation study in Table 7.2.

We see that the model trained from scratch achieves the worst performance with an mAP of 0.213. All methods using either supervised or self-supervised pretraining (AVCC) outperform this. For the supervised variants of the model, we

Ablation Study 2: AV Contrastive + Clustering (AVCC)			
Evaluation Dataset Labels		Audioset Balanced 527 sound classes	
Method	Pretraining	Size	mAP ($\uparrow$)
Scratch	None	0	0.213
Supervised	ImageNet	1.2M	0.219
Supervised	Kinetics 400	240k	0.225
Supervised	Kinetics 700	480k	0.231
Supervised	Audioset 240k	240k	0.239
AVCC	Kinetics 400	240k	0.221
AVCC	Kinetics 700	480k	0.224
AVCC	Audioset 240k	240k	0.243

Table 7.2: Ablation study for different types and sizes of pretraining dataset + type of pretraining (scratch, self-supervised (AVCC), fully supervised) for the audio-only backbone (2D Resnet-18) of all the methods. Our proposed pretraining on the Audioset dataset gives the best results compared to any other model variation.

see that the model trained on Audioset-240k is clearly the best with an mAP of 0.239. This points to the usage of an audio-specific curated dataset being beneficial to learning better audio features (which makes intuitive sense). The AVCC models trained on Kinetics-400 and Kinetics-700 outperform training from scratch and give an mAP of 0.221 and 0.224 respectively, which points to the fact that more pretraining data may help learn better representations. However, they both underperform the supervised variants trained on the same datasets (which get an mAP of 0.225 and 0.231 respectively). The best result we get is with the AVCC model pretrained on Audioset-240k, which gets an mAP of 0.243 and outperforms even its supervised counterpart which only gets an mAP of 0.239 (which is better than all other results). This highlights the usefulness of our pretraining procedure, especially when trained on a large-scale audio-specific dataset like Audioset.

Barlow Twins: Dimensionality of Features

The next ablation study we do is for the Barlow Twins method, where the dimensionality of the features extracted from the final projector MLP is an important parameter. The original Barlow Twins paper [222] shows that increasing this leads to consistent performance gains even for very large dimensionalities. We experiment with output dimensionalities from the set [256, 512, 1024, 2048, 4096]. We evaluate both the A-A Barlow Twins and the A-A + A-V Loc Barlow Twins

Ablation Study 3: Barlow Twins

Evaluation Dataset		Audioset Balanced				
Labels		527 sound classes				
Metric		mAP ($\uparrow$)				
Method	Pretraining	Output Feature Dimensionality				
		256	512	1024	2048	4096
A-A BT	Kinetics 400	0.219	0.228	0.235	0.238	0.237
A-A BT	Kinetics 700	0.220	0.226	0.236	0.235	0.237
A-A BT	Audioset 240k	0.233	0.241	0.249	0.248	0.248
A-A + A-V Loc BT	Kinetics 400	0.222	0.222	0.231	0.244	0.245
A-A + A-V Loc BT	Kinetics 700	0.224	0.228	0.244	0.246	0.245
A-A + A-V Loc BT	Audioset 240k	0.232	0.243	0.249	0.252	0.250

Table 7.3: Ablation study for feature dimensionality used for the Barlow Twins-based methods. We find that on average, a higher dimensionality helps but performance tends to be similar or saturates when increasing it after a point

methods pretrained on all the chosen pretraining datasets. Similar to the previous ablations, the chosen task is again audio event detection on balanced Audioset. We present results for this ablation study in Table 7.3. We find that the performance for each model variation increases initially with increase in dimensionality, but plateaus after a point. For the A-A BT method, the best dimensionality on Kinetics-400 is 2048 with an mAP of 0.238; on Kinetics-700 is 4096 with an mAP of 0.237; and on Audioset-240k is 1024 with an mAP of 0.249. For the A-A + A-V Loc BT method, the best dimensionality on Kinetics-400 is 4096 with an mAP of 0.245; on Kinetics-700 is 2048 with an mAP of 0.246; and on Audioset-240k is 2048 with an mAP of 0.252. All mAP values for higher dimensionalities are significantly higher than those for dimensionalities of 256 and 512. We also observe again that the models trained on the more audio specific Audioset yield an overall better performance irrespective of the chosen dimensionality and for both proposed model variations. For subsequent evaluation, we pick the model from this ablation study whichever gives the best performance (the chosen models for each configuration are indicated in bold in Table 7.3)

7.3.3 Downstream Audio Results

We now present downstream task results for 3 different audio recognition tasks: (a) environmental sound classification on ESC50 [156], (b) speaker identification on Voxceleb1 [135], and (c) audio event detection on the balanced set of Audioset

[69]. We evaluate the best performing variation of all the proposed methods on these downstream tasks. Unless stated otherwise, we finetune the pretrained audio encoder models on the downstream checkpoints. For the ESC50 and Voxceleb1 tasks, we use the cross-entropy loss for classification. For the Audioset task, we use class-wise binary cross-entropy because there are multiple classes present in each sample at different points of time. For all the tasks, we finetune for 50 epochs and use the Adam optimizer with a base learning rate of 0.0001 which is reduced by a factor of 10 at the 20th and 40th epochs (similar to the finetuning setup in the previous chapters). Results for downstream audio classification on ESC50 and Voxceleb1 can be found in Table 7.4, and those for audio event detection on balanced Audioset can be found in Table 7.5.

Downstream Audio Classification Results			
Evaluation Dataset		ESC50	VoxCeleb1
Labels		50 sound classes	1251 speaker IDs
Method	Pretraining	Accuracy ($\uparrow$)	Accuracy ($\uparrow$)
Scratch	None	0.543	0.738
AV Contrastive + Clustering	Kinetics 400	0.803	0.743
AV Contrastive + Clustering	Kinetics 700	0.829	0.742
AV Contrastive + Clustering	Audioset 240k	0.861	0.746
A-A SimSiam	Kinetics 400	0.824	0.762
A-V SimSiam	Audioset 240k	0.858	0.770
A-A Barlow Twins	Kinetics 400	0.821	0.769
A-A Barlow Twins	Audioset 240k	0.874	0.775
A-V Loc Barlow Twins	Kinetics 400	0.602	0.738
A-V Loc Barlow Twins	Audioset 240k	0.614	0.735
A-A + A-V Loc Barlow Twins	Kinetics 400	0.830	0.769
A-A + A-V Loc Barlow Twins	Audioset 240k	0.886	0.778
Supervised	ImageNet	0.823	0.751
Supervised	Audioset 240k	0.865	0.763

Table 7.4: Downstream results for audio classification tasks: (a) environmental sound classification on ESC50, (b) speaker identification on Voxceleb1. We use the pre-trained audio encoders from each method to do audio-only downstream recognition on both datasets.

We also compare with variations of the same audio encoder that are trained

from scratch from a random initialization and pretrained in a supervised manner on ImageNet. We find that training from scratch is by far the worst performing method on the tested datasets. The performance gap is especially huge on ESC50, where it only gets an accuracy of 0.543. This is likely due to the fact that ESC50 only has 2000 samples and is a small dataset, and pretraining is especially useful in evaluation settings like this. The gap between Scratch and other methods is smaller on VoxCeleb1, where it gets a fairly competitive accuracy of 0.738. This is likely because VoxCeleb1 has a fairly large training set where prior pretraining will still help, but produce diminishing returns when compared to a smaller dataset like ESC50.

The AVCC methods outperform training from scratch, with the variation trained on Audioset-240k being clearly the best out of the three with an ESC50 accuracy of 0.861 and a Voxceleb1 accuracy of 0.746. AVCC also outperforms both variations of the SimSiam methods on ESC50, but is worse than them on Voxceleb1. The results for the SimSiam methods are competitive with AVCC, but are worse than those obtained by the A-A Barlow Twins method, which provides very strong performance with an ESC50 accuracy of 0.874 and a Voxceleb1 accuracy of 0.775. This suggests that Barlow Twins may be a better self-supervised objective for audio-only representation learning on the tested datasets than SimSiam or a MoCo-style approach with AVCC. However, the results obtained with A-V Loc Barlow Twins are much worse than any other method, and the only baseline they beat is training from scratch. This could be due to the fact that it is difficult to learn a audiovisual localization map without any labels whatsoever and could lead to spurious correlations being learned between audio and video that may be useful in some cases but not generalizable broadly. This issue has also been studied in other works including that by Senocak et al. [176, 177], and current works try to alleviate this but are much more effective when supervision is used [132]. Note however, that even this form of pretraining produces some useful features, because the downstream results are better than those obtained by training from scratch.

When the two Barlow Twins based method (A-A and A-V Loc) are combined, this leads to our best overall model which clearly outperforms all other proposed methods on the chosen tasks when pretrained on Audioset-240k. The A-A + A-V Loc Barlow Twins method gets an ESC50 accuracy of 0.886 (which even outperforms supervised pretraining on both ImageNet and more interestingly on the same Audioset-240k too). This provides compelling evidence that the proposed form of joint audiovisual self-supervision is useful, and even though the A-V Barlow Twins method on its own may not yield the best features, it contains com-

plementary information to the A-A Barlow Twins method which can improve the performance of the combination by a nontrivial amount. Visual self-supervision for audio features seems promising even if it may not be the primary driver of the learning process.

Downstream Audio Event Detection Results			
	Evaluation Dataset	Audioset Balanced	
	Labels Modalities	527 sound classes, multiple per clip Audio-only	Audiovisual
Method	Pretraining	mAP ($\uparrow$)	mAP ($\uparrow$)
Scratch	None	0.213	0.249
GDT [148]	Kinetics 400	0.234	0.267
GDT [148]	Audioset 1.8M	0.264	0.305
AV Contrastive + Clustering	Kinetics 400	0.221	0.256
AV Contrastive + Clustering	Kinetics 700	0.224	0.257
AV Contrastive + Clustering	Audioset 240k	0.243	0.273
A-A SimSiam	Kinetics 400	0.232	0.269
A-V SimSiam	Audioset 240k	0.242	0.278
A-A Barlow Twins	Kinetics 400	0.238	0.276
A-A Barlow Twins	Audioset 240k	0.249	0.293
A-V Loc Barlow Twins	Kinetics 400	0.215	0.249
A-V Loc Barlow Twins	Audioset 240k	0.219	0.258
A-A + A-V Loc Barlow Twins	Kinetics 400	0.245	0.279
A-A + A-V Loc Barlow Twins	Audioset 240k	0.252	0.301
Supervised	ImageNet	0.219	0.253
Supervised	Audioset 240k	0.239	0.297

Table 7.5: Downstream results for audio event detection on the balanced Audioset. Note: when using audiovisual downstream evaluation for audio-only pretraining (e.g. for A-A BT and A-A SimSiam), the visual branch of the network is trained from scratch for downstream evaluation.

For downstream audio event detection on the balanced set of Audioset [69], we present results in Table 7.5. The balanced set of Audioset consists of around 22k train samples, which is roughly 1/10th the size of the 240k sized pretraining set that we use from the main training set of Audioset. There is no overlap between the pretraining set and the balanced training set. We present results for

two evaluation scenarios: (a) audio-only evaluation which is similar to the setting for all previous results discussed in this chapter, and (b) audiovisual evaluation which also uses the video backbones of the pretrained networks to extract video features that are concatenated to the output audio features before eventual classification. In cases where a video backbone is not available (e.g. A-A SimSiam and A-A Barlow Twins), the video network is trained from scratch for audiovisual evaluation while the pretrained audio network remains the same. In addition to all methods compared in the previous experiments, we also include results from GDT [148] pretrained on the full Audioset (with 1.8M samples as opposed to our subset of 240k). The presented results for Audioset Balanced follow a similar trend to those for ESC50. Training from scratch still produces the worst overall results (which is to be expected), which sets a floor of an audio-only mAP of 0.213 and audiovisual mAP of 0.249. The AVCC methods improve upon this, with the best performing variation once again being the one trained on Audioset-240k, which achieves an audio-only mAP of 0.243 and audiovisual mAP of 0.273. This is very comparable to the performance of A-V SimSiam with an audio-only mAP of 0.242, but a slightly better audiovisual mAP of 0.278. A-A Barlow Twins offers extremely strong performance, with an audio-only mAP of 0.249 and an audiovisual mAP of 0.293 even when the video branch is trained from scratch. Both of these are better than either AVCC or A-V SimSiam. A-V Loc Barlow Twins is still the second worst performing method on its own, but is able to provide complementary information to A-A Barlow Twins. This makes A-A + A-V Loc Barlow Twins the best overall method on the tested datasets with an audio-only mAP of 0.252 and audiovisual mAP of 0.301. The only baseline which outperforms this is the GDT version that uses 10x more training data with 1.8M total examples. However, because our method outperforms the Kinetics-400 variant of GDT with the exact same pretraining set, this shows definite promise that our audio representations may get even better with access to even more training data.

7.4 Discussion

Overall, the presented results offer strong evidence for audiovisual self-supervised learning being useful for learning general sound representation. A specific interesting part is the ability of visual part of joint audiovisual self-supervision to drive the learning of better audio representations even if the visual-only supervision may be individually suboptimal. This has been a recurrent theme in this

thesis (even for speech representation learning in previous chapters). Our proposed methods in some cases outperform even large scale supervised pretraining on the same dataset (Audioset), which strongly motivates the utility audiovisual self-supervision.

Out of the tested methods, the Barlow Twins based methods provided the best results. A possible explanation for this could be that the Barlow Twins methods are easier to train whereas the others are more brittle and prone to representation collapse (especially SimSiam). Combining the Barlow Twins objective with audiovisual sound source localization was an interesting idea with the intention of narrowing down onto the sound producing regions. However, if video encoding is done with newly proposed efficient video Transformer models which can process the video sequences as ‘tokens’ with a positional embedding, this may alleviate the need of explicitly learning localization because the cross-modal attention in these networks could automatically learn the most important relationships. Since these research has been conducted, there have been new works and architectures which may be a lot more appropriate and may yield even better results than those presented in this work for the same experimental setting.

Another interesting thing to note is that the AVCC work was conceptually very similar to the clustering prediction component of the currently state-of-the art AV-HuBerT [178] model but the key difference was not having a reliably good initialization (AV-HuBerT uses labels from MFCCs for the initial cluster assignments whereas ours were random). This was not something that we explored during the time the research for this work was being conducted, but makes intuitive sense to enforce a better starting point (initialization) for the audio representations.

Chapter 8

Conclusion

In this chapter, we summarize the core contributions of the thesis, along with highlighting certain limitations. We also talk about potential future directions and open questions in the area of this work. In this thesis, we have studied the utility of the visual self-supervision to guide the learning of better audio representations. Paired audiovisual data contains the unique theme of cross-modal correspondence and heterogeneity at the same time. Self-supervised learning can leverage these relationships and natural synchrony. For speech and audio data, the visual modality can offer a complementary and distinct supervisory signal that augments audio-only self-supervision even if it may not be individually optimal. The simultaneous abundance of unlabeled paired audiovisual data and paucity of labeled data for specific audio problems like emotion recognition and speech recognition with limited labeled data makes self-supervision a highly impactful area to make research contributions in. We studied audiovisual representation learning by using existing tools and methods in computer vision, machine learning, affective computing, speech and audio. Our proposed methods have been published in conferences and journals in affective computing, speech signal processing, computer vision, and machine learning.

8.1 Contributions

There are a number of strong contributions of the thesis. In this section, we highlight the most important ones:

Speech-driven facial animation as a pretext task

Despite speech-driven facial animation having been a long-studied problem, this work is the first work to study this as a pretext task for self-supervised learning of audio representations. The cross-modal correlation between **lip movements** and **speech content**, and that between **facial expressions** and **speech emotion** is particularly interesting, and downstream task performance on speech and emotion recognition is positively influenced by the proposed self-supervision.

Self-supervision on audio outperforms training from scratch

Our proposed methods for audio and audiovisual representation learning (across a variety of tasks and dataset sizes) outperform training from scratch by a good margin. This is an important result because it shows the usefulness of self-supervised pretraining (instead of just always training from scratch). Of particular note are the strong results obtained in limited labeled data settings (e.g. in Chapter 6), which is a crucial application area of self-supervised pretraining.

Joint audiovisual self-supervision is effective

In all experiments, the models using joint audiovisual self-supervision end up outperforming the models using either form of unimodal self-supervision. While this cannot always be taken for granted (see sections on potential limitations of multi-modal correlation based learning in Chapter 2), at least for this particular type of structured data (audiovisual speech with time aligned lip movements and speech content), joint audiovisual self-supervision is strong. This shows that both the redundancy and complementarity of both modalities are modeled effectively and each contribute their own gains.

Contrastive learning works for broad audio and audiovisual data

We have also shown in Chapter 7 that a combination of simple self-supervised learning methods like contrastive learning (combined with clustering), and simple siamese representation learning is effective for joint audiovisual representation learning. While this is not a super new contribution (there has been a good amount of concurrent and future work that also leverages this type of learning), it is a strong proof of such methods being applicable to large scale general audiovisual data.

Audiovisual sound source localization as a pretext task

A significant new contribution of the thesis is using self-supervised audiovisual sound source localization as a pretext task for audio and visual representation learning. Much like how speech-driven facial reconstruction uses the correlation between lip movements in video and speech content in audio, sound source localization can leverage the correlation between the pixels of the sound producing regions in the visual modality with the actual sounds themselves in the audio modality. Using this and combining it with the strong Barlow Twins method in a joint audiovisual manner yields strong results.

8.2 Limitations

Simpler pretext tasks at larger scales could be better

A crucial limitation of the proposed method (speech-driven facial reconstruction for audio representation learning) is that it requires a fairly specific type of data to work effectively. In the work, it has only been demonstrated on cropped and (nearly) frontal facing talking head videos. This is a tiny subset of the possible variations of audiovisual speech data. There has been newer work that improves the quality of speech-driven facial reconstruction (e.g. VASA from Microsoft), however the core problem here is the *highly specific* design of the pretext task.

If you contrast the proposed method to production systems for audiovisual representation learning deployed at companies like Meta (in products like Instagram Reels, or YouTube for videos), the methods are algorithmically *much simpler* (e.g. contrastive learning across modalities), but the data and infrastructure have been scaled up exponentially. Another parallel line of progress has been in language models, where the prediction task is extremely simple (predicting the next token), however scaling has led to this being a super effective method. A more direct comparison to audiovisual representation learning works like AV-Hubert [178] also shows a similar trend of a simpler idea being scaled more effectively ultimately being more production ready. If conducting this research from scratch again today, I would probably work on pretext tasks that are simpler and more scalable, but on larger datasets (e.g. thousands of hours of pretraining data instead of 40-100 hours).

Transformers / State Space Models could be better than CNNs/GRUs

While the work shows strong results for the time during which the research was conducted, the state-of-the-art has significantly improved since on the tested datasets due to architectural improvements (largely due to Transformers being applied effectively to the audio modality). Another extremely promising type of architecture are Selective State Space Models, which have shown to be extremely effective at recurrently modeling long sequences for modalities like audio. A combination of these architectural changes could have led to a huge quantitative jump in the benchmark results.

Data Augmentation

The scale of the data being relatively small (40-100 hours), data augmentations for audiovisual data could have helped. Visual augmentations like random horizontal flipping, some slight gaussian blur and whichever of the RandAugment augmentations make sense to apply while preserving the lip movements could have been useful. It can also help to use audio augmentations like SpecAugment, noise addition with a tuned schedule. Lastly, augmentations like MixUp and carefully chosen CutMix or Random Erasing could also help. However, these were not tried during the work.

Spurious Multimodal Correlations

As discussed in Chapter 2, we must be careful to not always assume that cross-modal patterns exist or are useful. We must also try to be mindful that our modeling approach does not learn spurious correlations in multimodal data. For example, in driving videos you will usually always see the sky in addition to the body of the car. Incorrectly modeling cross-modal correlations (e.g. for audio and video) could lead to the model learning to associate the car's engine sound with the visual content of the sky (which is clearly a spurious correlation). This is especially true when working on problems like self-supervised audiovisual sound source localization (in Chapter 7).

8.3 Possible Future Directions

A natural line of possible future work could stem from the limitations discussed in the previous section. Some of this work is fairly low-hanging fruit: e.g. using

better architectures like transformers/SSM's, performing audio and visual data augmentations, and scaling up to larger datasets. However, some of the questions are much more fundamental (e.g. how to correctly do cross-modal learning while avoiding spurious correlations). This requires more fundamental research about the science of multimodal learning (see [18]) and a deeper understanding of cross-modal interactions.

From a technical perspective, another compelling direction to investigate is the application of recent advances in Large Language Models (LLM's) to the audio and audiovisual domain. Treating an entire sequence of audiovisual data as tokens that go into a multimodal transformer, and incorporating a number of tricks that go into making larger and longer context models work at scale is a very high impact area. There has been some work that has shown this (e.g. the Gemini models by Google can ingest text, video, and audio together in a single inference pass), but there is a lot more work to be done here in the short term. Of particular note are better tokenization and input representation strategies for audio and audiovisual data.

Cross-modal learning has many other application areas which could benefit from similar types of self-supervision. Robotics, for example, has many shared properties with audiovisual speech data which this thesis leveraged. Similar to how we do lip-video reconstruction in this thesis, we can do RGB to Depth reconstruction, RGB to LiDAR prediction, proprioception sensing to occupancy prediction, and a number of other interesting cross-modal correlation based methods in a lot of robotics scenarios. These are all potentially promising avenues for robotics representation learning, which will be crucial to bringing the next wave of physical and embodied intelligence to our society.

Bibliography

- [1] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. IEEE transactions on pattern analysis and machine intelligence, 2018. 11
- [2] T. Afouras, J. S. Chung, and A. Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496, 2018. 12
- [3] T. Afouras, J. S. Chung, and A. Zisserman. Deep lip reading: a comparison of models and an online application. arXiv preprint arXiv:1806.06053, 2018. 11
- [4] T. Afouras, J. S. Chung, and A. Zisserman. Asr is all you need: Cross-modal distillation for lip reading. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2143–2147. IEEE, 2020. 11
- [5] 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. 23, 97, 102, 103, 104
- [6] J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelović, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, and A. Zisserman. Self-supervised multimodal versatile networks. arXiv preprint arXiv:2006.16228, 2020. 22
- [7] S. Albanie, G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox, and A. Zisserman. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16, pages 35–53. Springer, 2020. 11

-
- [8] H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, and D. Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33:9758–9770, 2020. [xiv](#), [21](#), [51](#), [81](#), [91](#), [92](#), [95](#), [96](#), [98](#), [99](#), [106](#), [107](#)
- [9] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182, 2016. [20](#)
- [10] R. Arandjelovic and A. Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017. [21](#)
- [11] R. Arandjelovic and A. Zisserman. Objects that sound. In *ECCV*, pages 435–451, 2018. [21](#), [44](#), [46](#), [58](#)
- [12] Y. Asano, M. Patrick, C. Rupprecht, and A. Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. *Advances in Neural Information Processing Systems*, 33:4660–4671, 2020. [22](#)
- [13] Y. M. Assael, B. Shillingford, S. Whiteson, and N. De Freitas. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*, 2016. [11](#)
- [14] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. *arXiv preprint arXiv:2301.08243*, 2023. [3](#), [16](#)
- [15] A. Baevski, H. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*, 2020. [19](#)
- [16] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022. [24](#)
- [17] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio. End-to-end attention-based large vocabulary speech recognition. In *2016*
-

- IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 4945–4949. IEEE, 2016. 11
- [18] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence, 41(2):423–443, 2018. 10, 120
- [19] M. K. Baskar, S. Watanabe, R. Astudillo, T. Hori, L. Burget, and J. Černocký. Self-supervised sequence-to-sequence asr using unpaired speech and text. arXiv preprint arXiv:1905.01152, 2019. 20
- [20] G. Bertasius, H. Wang, and L. Torresani. Is space-time attention all you need for video understanding? In ICML, volume 2, page 4, 2021. 16
- [21] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5:135–146, 2017. 18
- [22] C. Bregler, M. Covell, and M. Slaney. Video rewrite: Driving visual speech with audio. In Proceedings of the 24th annual conference on Computer graphics and interactive techniques, pages 353–360, 1997. 13, 40
- [23] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33: 1877–1901, 2020. 18
- [24] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4):335, 2008. 12, 33, 57
- [25] H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, and R. Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing, 5(4):377–390, 2014. 12, 33, 44, 57
- [26] M. Caron, P. Bojanowski, A. Joulin, and M. Douze. Deep clustering for unsupervised learning of visual features. In Proceedings of the European Conference on Computer Vision (ECCV), pages 132–149, 2018. 15, 95, 107

- [27] M. Caron, P. Bojanowski, J. Mairal, and A. Joulin. Leveraging large-scale uncurated data for unsupervised pre-training of visual features. arXiv preprint arXiv:1905.01278, 2019. 15
- [28] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Un-supervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882, 2020. 15, 16, 97
- [29] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pages 9650–9660, 2021. 16
- [30] D. M. Chan, S. Ghosh, D. Chakrabarty, and B. Hoffmeister. Multi-modal pre-training for automated speech recognition. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 246–250. IEEE, 2022. 22
- [31] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman. Soundspaces: Audio-visual navigation in 3d environments. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16, pages 17–36. Springer, 2020. 24
- [32] F. Chen, D. Zhang, M. Han, X. Chen, J. Shi, S. Xu, and B. Xu. Vlp: A survey on vision-language pre-training, 2022. 24
- [33] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman. Vggsound: A large-scale audio-visual dataset. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 721–725. IEEE, 2020. 24
- [34] H. Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, and A. Zisserman. Localizing visual sounds the hard way. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16867–16876, 2021. 23, 24
- [35] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu. Lip movements generation at a glance. In Proceedings of the European conference on computer vision (ECCV), pages 520–535, 2018. 11

- [36] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597–1607. PMLR, 2020. [xiv](#), [16](#), [91](#), [92](#), [95](#), [101](#)
- [37] X. Chen and K. He. Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 15750–15758, 2021. [16](#)
- [38] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020. [16](#)
- [39] Z. Chen, D. F. Fouhey, and A. Owens. Sound localization by self-supervised time delay estimation. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI, pages 489–508. Springer, 2022. [23](#)
- [40] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord. Unsupervised speech representation learning using wavenet autoencoders. IEEE/ACM transactions on audio, speech, and language processing, 27(12):2041–2053, 2019. [19](#)
- [41] J. Chung and A. Zisserman. Lip reading in the wild. In ACCV, 2016. [4](#), [21](#), [33](#), [44](#), [58](#), [82](#), [84](#)
- [42] J. S. Chung, A. Jamaludin, and A. Zisserman. You said that? arXiv preprint arXiv:1705.02966, 2017. [22](#)
- [43] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3444–3453. IEEE, 2017. [11](#), [12](#)
- [44] S.-W. Chung, H. G. Kang, and J. S. Chung. Seeing voices and hearing voices: learning discriminative embeddings using cross-modal self-supervision. arXiv preprint arXiv:2004.14326, 2020. [22](#)
- [45] Y. Chung, W. Hsu, H. Tang, and J. Glass. An unsupervised autoregressive model for speech representation learning. arXiv:1904.03240, 2019. [19](#), [29](#), [31](#), [32](#), [34](#), [35](#), [36](#), [45](#), [46](#), [53](#), [56](#), [58](#), [60](#), [61](#), [62](#), [82](#), [84](#)

- [46] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre. Detecting depression from facial actions and vocal prosody. In 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, pages 1–7. IEEE, 2009. 12
- [47] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America, 120(5):2421–2424, 2006. 43
- [48] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In Proceedings of the European Conference on Computer Vision (ECCV), pages 720–736, 2018. 26
- [49] V. R. de Sa. Learning classification with unlabeled data. Advances in neural information processing systems, pages 112–112, 1994. 14
- [50] B. Decoux. Self-supervised learning in cooperative stereo vision correspondence. International Journal of Neural Systems, 8(01):101–111, 1997. 14
- [51] M. Deitke, D. Batra, Y. Bisk, T. Campari, A. X. Chang, D. S. Chaplot, C. Chen, C. P. D’Arpino, K. Ehsani, A. Farhadi, et al. Retrospectives on the embodied ai workshop. arXiv preprint arXiv:2210.06849, 2022. 26
- [52] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009. 1, 4, 15
- [53] J. Devlin, M. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805, 2018. 14, 18
- [54] C. Doersch and A. Zisserman. Multi-task self-supervised visual learning. In Proceedings of the IEEE International Conference on Computer Vision, pages 2051–2060, 2017. 17
- [55] C. Doersch, A. Gupta, and A. Efros. Unsupervised visual representation learning by context prediction. In ICCV, pages 1422–1430, 2015. 15
- [56] J. Donahue and K. Simonyan. Large scale adversarial representation learning. In NeurIPS, pages 10541–10551, 2019. 15

- [57] P. Ekman et al. Basic emotions. Handbook of cognition and emotion, 98 (45-60):16, 1999. 12
- [58] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. arXiv:1804.03619, 2018. 22, 76
- [59] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In Proceedings of the 21st ACM international conference on Multimedia, pages 835–838, 2013. 34, 35, 36, 59, 60, 61, 62
- [60] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. ACM Transactions on Graphics (TOG), 21(3):388–398, 2002. 13, 41
- [61] H. M. Fayek and A. Kumar. Large scale audiovisual learning of sounds with weakly labeled data. In IJCAI, 2020. 19
- [62] A. Fernandez-Lopez and F. Sukno. Survey on automatic lip-reading in the era of deep learning. Image and Vision Computing, 2018. 11
- [63] B. Fernando, H. Bilen, E. Gavves, and S. Gould. Self-supervised video representation learning with odd-one-out networks. In CVPR, pages 3636–3645, 2017. xiii, 15, 29, 30, 31
- [64] J. W. Fisher III, T. Darrell, W. Freeman, and P. Viola. Learning joint statistical models for audio-visual fusion and segregation. Advances in neural information processing systems, 13, 2000. 23
- [65] R. Gao and K. Grauman. 2.5 d visual sound. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 324–333, 2019. 22
- [66] R. Gao and K. Grauman. Co-separating sounds of visual objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3879–3888, 2019. 23
- [67] R. Gao and K. Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 15490–15500. IEEE, 2021. 23

- [68] R. Gao, R. Feris, and K. Grauman. Learning to separate object sounds by watching unlabeled video. In Proceedings of the European Conference on Computer Vision (ECCV), pages 35–53, 2018. 23
- [69] 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, 2017. 26, 91, 92, 105, 106, 111, 113
- [70] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. arXiv:1803.07728, 2018. xiii, 15, 16, 17
- [71] S. Goldman and Y. Zhou. Enhancing supervised learning with unlabeled data. In ICML, pages 327–334. Citeseer, 2000. 14
- [72] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18995–19012, 2022. 26
- [73] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning, pages 369–376, 2006. 11
- [74] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems, 33:21271–21284, 2020. 16, 97
- [75] G. Hacohen and D. Weinshall. On the power of curriculum learning in training deep networks. arXiv preprint arXiv:1904.03626, 2019. 21
- [76] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic. Lips don’t lie: A generalisable and robust approach to face forgery detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5039–5049, 2021. 13

- [77] A. Haliassos, P. Ma, R. Mira, S. Petridis, and M. Pantic. Jointly learning visual and auditory speech representations from raw data. arXiv preprint arXiv:2212.06246, 2022. 22
- [78] A. Haliassos, R. Mira, S. Petridis, and M. Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14950–14962, 2022. 13
- [79] N. Harte and E. Gillen. Tcd-timit: An audio-visual corpus of continuous speech. IEEE Transactions on Multimedia, 17(5):603–615, 2015. 44
- [80] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 1, 81, 99, 102
- [81] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In CVPR, pages 9729–9738, 2020. xiv, 15, 16, 22, 91, 93, 94, 95, 97, 100
- [82] K. He, H. Fan, Y. Wu, S. Xie, and R. 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. xiv, 16, 93, 94
- [83] 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, pages 16000–16009, 2022. 16
- [84] J. Hershey and J. Movellan. Audio vision: Using audio-visual synchrony to locate sounds. Advances in neural information processing systems, 12, 1999. 23
- [85] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. 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. 19
- [86] 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. 19

- [87] D. Hu, F. Nie, and X. Li. Deep multimodal clustering for unsupervised audiovisual learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9248–9257, 2019. 21
- [88] D. Hu, R. Qian, M. Jiang, X. Tan, S. Wen, E. Ding, W. Lin, and D. Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. Advances in Neural Information Processing Systems, 33:10077–10087, 2020. 23
- [89] J. Huh, J. Chalk, E. Kazakos, D. Damen, and A. Zisserman. Epic-sounds: A large-scale dataset of actions that sound. arXiv preprint arXiv:2302.00646, 2023. 26
- [90] 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, pages 378–386, 2020. 13
- [91] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning, pages 448–456. pmlr, 2015. 15
- [92] H. Izadinia, I. Saleemi, and M. Shah. Multimodal analysis for identification and segmentation of moving-sounding objects. IEEE Transactions on Multimedia, 15(2):378–390, 2012. 23
- [93] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. arXiv preprint arXiv:2107.14795, 2021. 24
- [94] L. Jing, X. Yang, J. Liu, and Y. Tian. Self-supervised spatiotemporal feature learning via video rotation prediction. arXiv preprint arXiv:1811.11387, 2018. 16
- [95] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017. 4, 105
- [96] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen. Slow-fast auditory streams for audio recognition. arXiv preprint arXiv:2103.03516, 2021. 20

- [97] E. Kharitonov, M. Rivière, G. Synnaeve, L. Wolf, P.-E. Mazaré, M. Douze, and E. Dupoux. Data augmenting contrastive learning of speech representations in the time domain. In 2021 IEEE Spoken Language Technology Workshop (SLT), pages 215–222. IEEE, 2021. 19
- [98] E. Kidron, Y. Y. Schechner, and M. Elad. Pixels that sound. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), volume 1, pages 88–95. IEEE, 2005. 23
- [99] D. Kim, D. Cho, D. Yoo, and I. S. Kweon. Learning image representations by completing damaged jigsaw puzzles. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 793–802. IEEE, 2018. 17
- [100] A. Kolesnikov, X. Zhai, and L. Beyer. Revisiting self-supervised visual representation learning. arXiv preprint arXiv:1901.09005, 2019. 16
- [101] D. Kollias, S. Cheng, E. Ververas, I. Kotsia, and S. Zafeiriou. Generating faces for affect analysis. arXiv preprint arXiv:1811.05027, 2018. 13
- [102] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. International Journal of Computer Vision, pages 1–23, 2019. 12
- [103] B. Korbar, D. Tran, and L. Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In NeurIPS, pages 7763–7774, 2018. 21, 44, 46, 58
- [104] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, B. Schuller, K. Star, et al. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. arXiv preprint arXiv:1901.02839, 2019. 12, 33, 57
- [105] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84–90, 2017. 1
- [106] A. Kumar and V. Ithapu. A sequential self teaching approach for improving generalization in sound event recognition. In International Conference on Machine Learning, pages 5447–5457. PMLR, 2020. 19

- [107] A. Kumar and V. K. Ithapu. Secost:: Sequential co-supervision for large scale weakly labeled audio event detection. In *ICASSP 2020*, pages 666–670. IEEE, 2020. 19
- [108] Y. LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62, 2022. 15
- [109] 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. 16
- [110] J. Lian, A. Baevski, W.-N. Hsu, and M. Auli. Av-data2vec: Self-supervised learning of audio-visual speech representations with contextualized target representations. *arXiv preprint arXiv:2302.06419*, 2023. 22
- [111] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 4
- [112] S. Ling and Y. Liu. Decoar 2.0: Deep contextualized acoustic representations with vector quantization. *arXiv preprint arXiv:2012.06659*, 2020. 19
- [113] X. Liu, R. Qian, H. Zhou, D. Hu, W. Lin, Z. Liu, B. Zhou, and X. Zhou. Visual sound localization in the wild by cross-modal interference erasing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1801–1809, 2022. 23
- [114] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pre-training approach. *arXiv preprint arXiv:1907.11692*, 2019. 18
- [115] S. Livingstone and F. 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. 12, 33, 44, 57
- [116] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 24

- [117] P. Ma, R. Mira, S. Petridis, B. W. Schuller, and M. Pantic. Lira: Learning visual speech representations from audio through self-supervision. *arXiv preprint arXiv:2106.09171*, 2021. 22
- [118] P. Ma, S. Petridis, and M. Pantic. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7613–7617. IEEE, 2021. 11
- [119] P. Ma, S. Petridis, and M. Pantic. Visual speech recognition for multiple languages in the wild. *Nature Machine Intelligence*, pages 1–10, 2022. 11
- [120] S. Ma, Z. Zeng, D. McDuff, and Y. Song. Learning audio-visual representations with active contrastive coding. *arXiv preprint arXiv:2009.09805*, 2020. 22, 77
- [121] S. Ma, Z. Zeng, D. McDuff, and Y. Song. Contrastive learning of global and local video representations. *Advances in Neural Information Processing Systems*, 34:7025–7040, 2021. 22
- [122] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria. Multi-modal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems*, 161:124–133, 2018. 33, 57
- [123] T. Makino, H. Liao, Y. Assael, B. Shillingford, B. Garcia, O. Braga, and O. Siohan. Recurrent neural network transducer for audio-visual speech recognition. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 905–912. IEEE, 2019. 11
- [124] Y. Masuyama, Y. Bando, K. Yatabe, Y. Sasaki, M. Onishi, and Y. Oikawa. Self-supervised neural audio-visual sound source localization via probabilistic spatial modeling. *arXiv preprint arXiv:2007.13976*, 2020. 23
- [125] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264 (5588):746, 1976. 6, 10, 11
- [126] A. N. Meltzoff and M. K. Moore. Imitation of facial and manual gestures by human neonates. *Science*, 198(4312):75–78, 1977. 11
- [127] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 18

-
- [128] I. Misra. Visual Learning with Minimal Human Supervision. PhD thesis, 8 2018. doi: 10.1184/R1/7189397.v1. URL https://kilthub.cmu.edu/articles/thesis/Visual_Learning_with_Minimal_Human_Supervision/7189397. 3
- [129] I. Misra and L. v. d. Maaten. Self-supervised learning of pretext-invariant representations. In CVPR, pages 6707–6717, 2020. 15, 16
- [130] I. Misra, C. L. Zitnick, and M. Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In European Conference on Computer Vision, pages 527–544. Springer, 2016. xiii, 16, 18
- [131] S. Mo and P. Morgado. A closer look at weakly-supervised audio-visual source localization. arXiv preprint arXiv:2209.09634, 2022. 24
- [132] S. Mo and P. Morgado. Localizing visual sounds the easy way. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII, pages 218–234. Springer, 2022. 24, 112
- [133] P. Morgado, N. Vasconcelos, T. Langlois, and O. Wang. Self-supervised generation of spatial audio for 360 video. arXiv preprint arXiv:1809.02587, 2018. 22
- [134] 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, pages 12475–12486, 2021. 22
- [135] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612, 2017. 92, 106, 110
- [136] A. Nagrani, S. Albanie, and A. Zisserman. Learnable pins: Cross-modal embeddings for person identity. In Proceedings of the European Conference on Computer Vision (ECCV), pages 71–88, 2018. 22
- [137] T.-S. Nguyen, S. Stueker, and A. Waibel. Learning shared encoding representation for end-to-end speech recognition models. arXiv preprint arXiv:1904.02147, 2019. 19
-

- [138] M. Noroozi, A. Vinjimoor, P. Favaro, and H. Pirsiavash. Boosting self-supervised learning via knowledge transfer. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 9359–9367, 2018. 17
- [139] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik. Speech2face: Learning the face behind a voice. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7539–7548, 2019. 22
- [140] A. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018. 19, 28, 34, 35, 36, 45, 46, 47, 58, 60, 61, 62
- [141] 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. 21
- [142] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Ambient sound provides supervision for visual learning. In European conference on computer vision, pages 801–816. Springer, 2016. 22
- [143] A. Owens, J. Wu, J. McDermott, W. Freeman, and A. Torralba. Learning sight from sound: Ambient sound provides supervision for visual learning. IJCV, 126(10):1120–1137, 2018. 21, 22
- [144] S. Pal, A. Datta, and D. D. Majumder. Computer recognition of vowel sounds using a self-supervised learning algorithm. J. Anatomical Soc. India, 6:117–123, 1978. 14
- [145] X. Pan, P. Chen, Y. Gong, H. Zhou, X. Wang, and Z. Lin. Leveraging unimodal self-supervised learning for multimodal audio-visual speech recognition. arXiv preprint arXiv:2203.07996, 2022. 22
- [146] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio. Learning problem-agnostic speech representations from multiple self-supervised tasks. arXiv:1904.03416, 2019. xiii, 19, 20, 29, 34, 35, 36, 45, 46, 47, 58, 60, 61, 62, 82, 83, 84
- [147] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and

- S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems 32, pages 8024–8035, 2019. 32, 56
- [148] M. Patrick, Y. M. Asano, R. Fong, J. F. Henriques, G. Zweig, and A. Vedaldi. Multi-modal self-supervision from generalized data transformations. arXiv preprint arXiv:2003.04298, 2020. xiv, 13, 22, 77, 81, 91, 92, 93, 95, 98, 99, 101, 102, 105, 106, 113, 114
- [149] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. arXiv:1802.05365, 2018. 18
- [150] S. Petridis and M. Pantic. Prediction-based audiovisual fusion for classification of non-linguistic vocalisations. IEEE Transactions on Affective Computing, 7(1):45–58, 2015. 12, 21
- [151] S. Petridis, J. Shen, D. Cetin, and M. Pantic. Visual-only recognition of normal, whispered and silent speech. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6219–6223. IEEE, 2018. 11
- [152] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic. End-to-end audiovisual speech recognition. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6548–6552. IEEE, 2018. 10, 11
- [153] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic. Audiovisual speech recognition with a hybrid ctc/attention architecture. In 2018 IEEE Spoken Language Technology Workshop (SLT), pages 513–520. IEEE, 2018. 11
- [154] S. Petridis, Y. Wang, P. Ma, Z. Li, and M. Pantic. End-to-end visual speech recognition for small-scale datasets. arXiv preprint arXiv:1904.01954, 2019. 12
- [155] H. Pham, P. Liang, T. Manzini, L. Morency, and B. Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In AAAI, volume 33, pages 6892–6899, 2019. 21

-
- [156] K. J. Piczak. Esc: Dataset for environmental sound classification. In Proceedings of the 23rd ACM international conference on Multimedia, pages 1015–1018, 2015. [26](#), [92](#), [106](#), [110](#)
- [157] A. Piergiovanni, A. Angelova, and M. S. Ryoo. Evolving losses for unsupervised video representation learning. CVPR, 2020. [16](#), [21](#)
- [158] K. Prajwal, T. Afouras, and A. Zisserman. Sub-word level lip reading with visual attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5162–5172, 2022. [12](#)
- [159] V. Pratap, A. Hannun, Q. Xu, J. Cai, J. Kahn, G. Synnaeve, V. Liptchinsky, and R. Collobert. Wav2letter++: A fast open-source speech recognition system. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6460–6464. IEEE, 2019. [19](#)
- [160] R. Qian, D. Hu, H. Dinkel, M. Wu, N. Xu, and W. Lin. Multiple sound sources localization from coarse to fine. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16, pages 292–308. Springer, 2020. [23](#)
- [161] F. d. C. Quitry, M. Tagliasacchi, and D. Roblek. Learning audio representations via phase prediction. arXiv preprint arXiv:1910.11910, 2019. [19](#)
- [162] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019. [18](#)
- [163] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 2022. [13](#)
- [164] M. Ravanelli and Y. Bengio. Interpretable convolutional filters with sincnet. IEEE SLT Workshop, 2018. [19](#), [76](#)
- [165] M. Ravanelli and Y. Bengio. Learning speaker representations with mutual information. arXiv:1812.00271, 2018. [19](#)
- [166] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio. Light gated recurrent units for speech recognition. IEEE Transactions on Emerging Topics in Computational Intelligence, 2(2):92–102, 2018. [19](#), [37](#), [63](#), [76](#), [85](#)
-

- [167] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio. Multi-task self-supervised learning for robust speech recognition. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6989–6993. IEEE, 2020. 19
- [168] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), pages 1–8. IEEE, 2013. 12, 33, 57
- [169] M. Rivière, A. Joulin, P.-E. Mazaré, and E. Dupoux. Unsupervised pretraining transfers well across languages. ICASSP, 2020. 19, 29
- [170] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In MICCAI, pages 234–241, 2015. 42, 52
- [171] A. Rouditchenko, H. Zhao, C. Gan, J. McDermott, and A. Torralba. Self-supervised audio-visual co-segmentation. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2357–2361. IEEE, 2019. 21
- [172] J. A. Russell. A circumplex model of affect. Journal of personality and social psychology, 39(6):1161, 1980. 12
- [173] J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In Proceedings of the 22nd ACM international conference on Multimedia, pages 1041–1044, 2014. 26
- [174] J. Schmidhuber. Making the world differentiable: On using self-supervised fully recurrent neural networks for dynamic reinforcement learning and planning in non-stationary environments. Technical Report, 1990. 14
- [175] S. Schneider, A. Baevski, R. Collobert, and M. Auli. wav2vec: Unsupervised pre-training for speech recognition. arXiv:1904.05862, 2019. 19, 29, 82, 84
- [176] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. S. Kweon. Learning to localize sound source in visual scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4358–4366, 2018. 23, 112

- [177] A. Senocak, H. Ryu, J. Kim, and I. S. Kweon. Learning sound localization better from semantically similar samples. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4863–4867. IEEE, 2022. 23, 112
- [178] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. arXiv preprint arXiv:2201.02184, 2022. 22, 115, 118
- [179] B. Shillingford, Y. Assael, M. W. Hoffman, T. Paine, C. Hughes, U. Prabhu, H. Liao, H. Sak, K. Rao, L. Bennett, et al. Large-scale visual speech recognition. arXiv preprint arXiv:1807.05162, 2018. 11
- [180] A. Shukla. Learning self-supervised multimodal representations of human behaviour. In Proceedings of the 28th ACM International Conference on Multimedia, pages 4748–4751, 2020. 9
- [181] A. Shukla, S. S. Gullapuram, H. Katti, K. Yadati, M. Kankanhalli, and R. Subramanian. Affect recognition in ads with application to computational advertising. In Proceedings of the 25th ACM international conference on Multimedia, pages 1148–1156, 2017. 12
- [182] A. Shukla, S. S. Gullapuram, H. Katti, K. Yadati, M. Kankanhalli, and R. Subramanian. Evaluating content-centric vs. user-centric ad affect recognition. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, pages 402–410, 2017. 12
- [183] A. Shukla, H. Katti, M. Kankanhalli, and R. Subramanian. Looking beyond a clever narrative: Visual context and attention are primary drivers of affect in video advertisements. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, pages 210–219, 2018. 12
- [184] A. Shukla, S. S. Gullapuram, H. Katti, M. Kankanhalli, S. Winkler, and R. Subramanian. Recognition of advertisement emotions with application to computational advertising. IEEE Transactions on Affective Computing, 2020. 12
- [185] A. Shukla, S. Petridis, and M. Pantic. Visual self-supervision by facial reconstruction for speech representation learning. Sight and Sound Workshop, CVPR, 2020. 8, 9, 52, 82

- [186] A. Shukla, S. Petridis, and M. Pantic. Learning speech representations from raw audio by joint audiovisual self-supervision. Workshop on Self-supervision in Audio and Speech, ICML, 2020. 9, 52, 79
- [187] A. Shukla, K. Vougioukas, P. Ma, S. Petridis, and M. Pantic. Visually guided self supervised learning of speech representations. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6299–6303. IEEE, 2020. 7, 8, 9, 40, 52, 58, 82
- [188] A. Shukla, S. Petridis, and M. Pantic. Does visual self-supervision improve learning of speech representations for emotion recognition. IEEE Transactions on Affective Computing, 2021. 7, 8, 9, 27, 50, 82, 84
- [189] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014. 104
- [190] L. Smaira, J. Carreira, E. Noland, E. Clancy, A. Wu, and A. Zisserman. A short note on the kinetics-700-2020 human action dataset. arXiv preprint arXiv:2010.10864, 2020. 105
- [191] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley. Detection and classification of acoustic scenes and events. IEEE Transactions on Multimedia, 17(10):1733–1746, 2015. 20
- [192] W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. The journal of the acoustical society of america, 26(2):212–215, 1954. 11
- [193] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 8934–8943, 2018. 16
- [194] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (TOG), 36(4):1–13, 2017. 13, 41
- [195] M. Tagliasacchi, B. Gfeller, F. Quitry, and D. Roblek. Self-supervised audio representation learning for mobile devices. arXiv:1905.11796, 2019. 20
- [196] M. Tan and Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946, 2019. 15

- [197] Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. arXiv preprint arXiv:1906.05849, 2019. 21
- [198] Y. Tian, D. Li, and C. Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, pages 436–454. Springer, 2020. 23
- [199] G. Towell. Using unlabeled data for supervised learning. Advances in neural information processing systems, 8, 1995. 14
- [200] 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, pages 6450–6459, 2018. 99, 102
- [201] E. Tzinis, S. Wisdom, A. Jansen, S. Hershey, T. Remez, D. P. Ellis, and J. R. Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. arXiv preprint arXiv:2011.01143, 2020. 23
- [202] E. Tzinis, S. Wisdom, T. Remez, and J. R. Hershey. Audioscopev2: Audio-visual attention architectures for calibrated open-domain on-screen sound separation. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII, pages 368–385. Springer, 2022. 23
- [203] A. Varga and H. J. Steeneken. Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech communication, 12(3):247–251, 1993. 72, 86
- [204] 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. 1, 16
- [205] C. Vondrick, A. Shrivastava, A. Fathi, S. Guadarrama, and K. Murphy. Tracking emerges by colorizing videos. In Proceedings of the European Conference on Computer Vision (ECCV), pages 391–408, 2018. xiii, 16, 17
- [206] K. Vougioukas, S. Petridis, and M. Pantic. End-to-end speech-driven facial animation with temporal gans. arXiv:1805.09313, 2018. 13, 41, 52, 54

- [207] K. Vougioukas, P. Ma, S. Petridis, and M. Pantic. Video-driven speech reconstruction using generative adversarial networks. *Proc. Interspeech 2019*, pages 4125–4129, 2019. 13, 77
- [208] K. Vougioukas, S. Petridis, and M. Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, pages 1–16, 2019. 13, 53, 54
- [209] 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*, pages 12695–12705, 2020. 21
- [210] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2015. 16
- [211] X. Wang, A. Jabri, and A. A. Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019. 15, 16
- [212] P. Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018. 33, 44, 58, 82, 84
- [213] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai. Espnet: End-to-end speech processing toolkit. In *Interspeech*, 2018. 46
- [214] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 16
- [215] D. Wei, J. J. Lim, A. Zisserman, and W. T. Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8060, 2018. 16, 30
- [216] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le. Self-training with noisy student improves imagenet classification. In *CVPR*, pages 10687–10698, 2020. 15
- [217] B. Xu, C. Lu, Y. Guo, and J. Wang. Discriminative multi-modality speech recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14433–14442, 2020. 11

- [218] Q. Xu, A. Baevski, T. Likhomanenko, P. Tomasello, A. Conneau, R. Collobert, G. Synnaeve, and M. Auli. Self-training and pre-training are complementary for speech recognition. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3030–3034. IEEE, 2021. 20
- [219] K. Yamauchi, M. Oota, and N. Ishii. A self-supervised learning system for pattern recognition by sensory integration. Neural Networks, 12(10):1347–1358, 1999. 14
- [220] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv preprint arXiv:1606.06259, 2016. 12
- [221] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency. Multi-modal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2236–2246, 2018. 12
- [222] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction. In International Conference on Machine Learning, pages 12310–12320. PMLR, 2021. xiv, 16, 91, 92, 97, 102, 103, 109
- [223] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer. S4l: Self-supervised semi-supervised learning. In ICCV, pages 1476–1485, 2019. 16
- [224] J. Zhang, Z. Yin, P. Chen, and S. Nichele. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. Information Fusion, 59:103–126, 2020. 12
- [225] Z. Zhang, B. Wu, and B. Schuller. Attention-augmented end-to-end multi-task learning for emotion prediction from speech. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6705–6709. IEEE, 2019. 12
- [226] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba. The sound of pixels. In Proceedings of the European conference on computer vision (ECCV), pages 570–586, 2018. 23

-
- [227] H. Zhao, C. Gan, W.-C. Ma, and A. Torralba. The sound of motions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1735–1744, 2019. 23
- [228] H. Zhu, M. Luo, R. Wang, A. Zheng, and R. He. Deep audio-visual learning: A survey. arXiv preprint arXiv:2001.04758, 2020. 13, 22