INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

LiRA: Learning Visual Speech Representations from Audio through Self-supervision

Pingchuan Ma[1*], *Rodrigo Mira*[1*], *Stavros Petridis*[2], *Björn W. Schuller*[1,3] *and Maja Pantic*[1,2]

[1]iBUG Group, Imperial College London, UK
[2]Facebook London, UK
[3]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
pm4115@ic.ac.uk, rs2517@ic.ac.uk

Abstract

The large amount of audiovisual content being shared online today has drawn substantial attention to the prospect of audio-visual self-supervised learning. Recent works have focused on each of these modalities separately, while others have attempted to model both simultaneously in a cross-modal fashion. However, comparatively little attention has been given to leveraging one modality as a training objective to learn from the other. In this work, we propose Learning visual speech Representations from Audio via self-supervision (LiRA). Specifically, we train a ResNet+Conformer model to predict acoustic features from unlabelled visual speech. We find that this pre-trained model can be leveraged towards word-level and sentence-level lip-reading through feature extraction and fine-tuning experiments. We show that our approach significantly outperforms other self-supervised methods on the Lip Reading in the Wild (LRW) dataset and achieves state-of-the-art performance on Lip Reading Sentences 2 (LRS2) using only a fraction of the total labelled data.

Index Terms: self-supervised learning, lip-reading, visual speech recognition, visual representations, conformer

1. Introduction

Self-supervised learning aims to leverage unlabelled data by extracting the training objective directly from the input itself, in an attempt to model meaningful representations of the proposed modality which capture its content and structure. In works adopting this methodology, this task is usually known as the "pretext task" and this initial training procedure is known as the "pre-training" stage. After pre-training, the network is trained on the "downstream task", which generally involves a smaller set of manually labelled data. This methodology has received substantial attention in recent years within the computer vision community. Pretext tasks for visual self-supervision include image colourisation [42], jigsaw puzzle solving [21], as well as combinations of these and other tasks [12]. Self-supervised learning has also been explored in the speech community through works such as Contrastive Predicting Coding (CPC) [22] and wav2vec [34], which predict/discriminate future segments of audio samples; LIM (Local Info Max) [32], which maximises mutual information for the same speaker; and, more recently, PASE (Problem Agnostic Speech Encoder) [26, 33], which predicts established audio features such as STFT and MFCC.

Self-supervision has also been adopted in the audiovisual domain. Recent approaches include audiovisual fusion [27, 28], clustering [4], and distillation [31]; cross-modal discrimination [23]; cyclic translation between modalities [30]; and permutative predictive coding [38]. Shukla *et al.* [35] focus on learning audio representations by facial reconstruction from waveform speech. Conversely, [24] predict frequency-based summaries of ambient sound from video, while other recent works apply audio-visual synchronisation [5, 7, 13] to learn visual embeddings. A task that can benefit from self-supervised learning is lip-reading. Current state-of-the-art lip-reading models rely on annotating hundreds of hours of visual speech data [18], which is costly. To solve this issue, Afouras *et al.* [3] propose using a pre-trained Automatic Speech Recognition (ASR) model to produce machine-generated captions for unsupervised pre-training. This provides automatically labelled data but still relies on an ASR model trained on large amounts of labelled data.

In this work, we aim to leverage the vast amount of available audiovisual speech data to learn generic visual speech features and improve state-of-the-art lip-reading models by predicting audio features from visual speech. The targeted audio features are extracted from waveform audio without the need for additional labels using an established speech encoder (PASE+ [33]). Using the proposed approach, the learnt visual features are explicitly guided by audio which contains rich information about speech. This in turn can lead to learning visual features which are more suitable for speech recognition. After this training procedure, we apply our model (Fig. 2) for lip-reading on a transcribed visual speech dataset.

Our research contributions are as follows: **1)** We present LiRA, which learns powerful visual speech representations by predicting acoustic features from raw video taken from large audio-visual datasets. **2)** We demonstrate that LiRA provides a good initialisation for fine-tuning lip-reading models which consistently outperforms training from scratch, and that this method is particularly beneficial for smaller labelled datasets. **3)** We show that LiRA outperforms previous self-supervised methods for word-level lip-reading, achieving an accuracy of 88.1% on LRW by pre-training on unlabelled data. **4)** Finally, we leverage our self-supervised approach towards sentence-level lip-reading, and find that our fine-tuned model achieves state-of-the-art performance for LRS2.

2. Methodology

2.1. Pretext task

LiRA predicts PASE+ features from raw video and is composed of three distinct components. The first is the spatial encoder, which is a traditional 2D ResNet-18 preceded by a 3D front-end layer. The second component is the temporal encoder – the conformer – which receives as input the frame-wise features produced by the spatial encoder and returns a set of features of the same size. The conformer encoder combines traditional attention-based transformer blocks, which excel at capturing global temporal dependencies, with convolutional layers, which

[*] equal contribution

http://dx.doi.org/10.21437/Interspeech.2021-1360

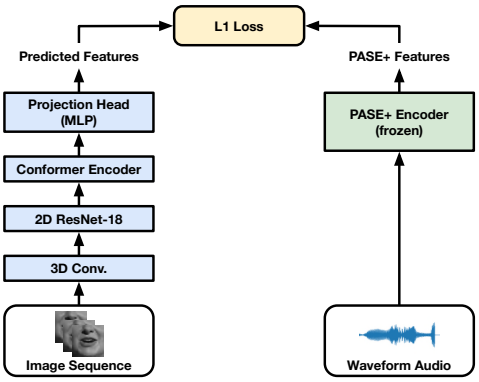


Figure 1: *The high-level architecture of our model and our methodology for audiovisual self-supervised training.*

model local patterns efficiently [10]. The final component is the projection head (based on the MLP – Multi-Layer Perceptron – workers presented in [26]), which projects these representations into the predicted PASE+ features. To train the model, we apply an L1 loss between the generated embeddings and the features extracted from the pre-trained (frozen) PASE+ model, as shown in Fig. 1. We would also like to mention that we have also experimented with predicting MFCC features but the results were worse than predicting PASE+ features.

2.2. Downstream task

To evaluate the visual speech representations, we run three variations of end-to-end lip-reading experiments. The training procedure is illustrated in Fig. 2. LiRA-Supervised models are trained from scratch based on the same encoder as in the self-supervised training [16]. This serves as our baseline model since it is trained only with the labelled training data. LiRA-Frozen models are trained using LiRA features from the pre-trained encoder. This allows us to evaluate the visual representations learned during self-supervised learning. Finally, LiRA-FineTuned models use the same model as LiRA-Supervised but are initialised with the pre-trained encoder weights from the pretext task. By using this configuration, we can evaluate the model initialisation capabilities of the proposed self-supervised learning approach. For each of these methods, we adopt a separate model for each lip-reading task - six models in total. For word-level lip-reading, we use a Multi-Scale Temporal Convolutional Network (MS-TCN) [19] on top of the encoder, followed by a linear classifier for classification. For sentence-level lip-reading, we follow the state-of-the-art lip-reading model [16] on LRS2 and build a hybrid CTC/attention model. We use the same conformer encoder architecture as in the pre-training phase, followed by the transformer decoder for sequence-to-sequence training [39]. We also perform fine-tuning experiments using the pre-trained model.

3. Experimental Setup

3.1. Datasets

In this work, we use an unlabelled version of Lip Reading Sentences 3 (LRS3) for pre-training and evaluate the performance of speech representations on LRW and LRS2. LRW [6] is comprised of approximately 500 000 1.16 second labelled utterances (173 hours in total) featuring a specific word from a 500 word vocabulary. It features hundreds of different speakers recorded in a variety of different backgrounds and head poses. LRS2 [1] is composed of approximately 150 000 transcribed utterances of varying lengths (224.5 hours in total). This corpus presents

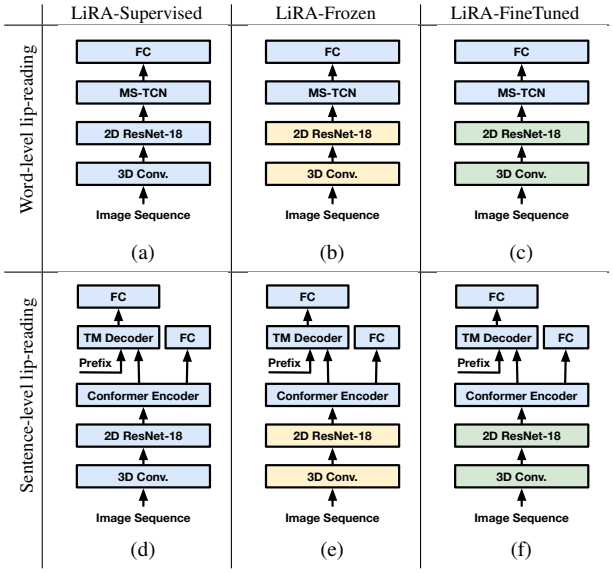


Figure 2: *The variations of the end-to-end lip-reading architecture. The sub-figures in the top row ((a),(b),(c)) refer to the word-level lip-reading training procedures, while the sub-figures in the bottom row ((d),(e),(f)) refer to sentence-level lip-reading. From left to right, (a) and (d) denote training from scratch (the whole model is initialised randomly); (b) and (e) are feature extraction experiments based on visual features extracted from the pre-trained model; and (c) and (f) are fine-tuning experiments. Blue coloured blocks are trained from scratch on the downstream task; yellow coloured blocks are loaded from the pre-trained model and kept frozen during the downstream task; and green coloured blocks are loaded from the pre-trained model and are then fine-tuned for the downstream task. We abbreviate the following model layers: TM: Transformer, FC: Fully-Connected layer, MS-TCN: Multi-Scale Temporal Convolutional Network.*

a greater challenge since it features a largely unconstrained vocabulary of more than 40 000 words. Both datasets are collected from BBC programs.

LRS3 [2] similarly contains approximately 150 000 utterances of varying lengths (438.9 hours in total) taken from TED talks. However, these utterances are substantially longer than the ones featured in LRS2, resulting in effectively double the total amount of hours of video and a larger vocabulary. This dataset guarantees no overlap between the speakers featured in the train and test sets, meaning that the test set is entirely comprised of speakers that were not seen in other sets.

3.2. Pre-processing

To crop the mouth Regions of Interest (ROIs), we start by detecting the 68-point facial landmarks using dlib [11]. We then normalise each frame using a neutral reference frame to remove rotation and size differences. Given the transformed facial landmarks, a fixed bounding box is used to crop mouth ROIs with a size of 96×96.

3.3. Data augmentation

Following [15], we produce augmented visual streams by applying the techniques of horizontal flipping with a probability of 0.5 and random cropping to a size of 88×88. During the testing phase, instead of randomly cropping, we crop a patch of size 88×88 from the centre of the image. For the word-level classification, mixup with a weight of 0.4 is employed.

Table 1: *A comparison of the performance between the baseline methods and ours (pre-trained on LRS3) on the LRW dataset.*

Methods	Strategy	Acc. (%)
ResNet + BLSTM [37]	Supervised	83.0
Two-stream 3D CNN [40]	Supervised	84.1
ResNet + BLSTM [36]	Supervised	84.3
ResNet + DenseTCN [17]	Supervised	88.4
PerfectMatch [7]	Self-supervised	71.6
PT-CDDL [8]	Self-supervised	75.9
AV-PPC [38]	Self-supervised	84.8
LiRA-Supervised [15]	Supervised	87.4
LiRA-Frozen	Self-supervised	83.1
LiRA-FineTuned	Self-supervised	88.1

3.4. Training settings in the pretext task

The 3D front-end module preceding our ResNet consists of a convolutional layer with kernel size (5, 7, 7) followed by a max pooling layer. The conformer, on the other hand, is comprised of an initial embedding module – feed forward layer combined with layer normalisation, dropout (0.1), activation (ReLU – Rectified Linear Unit) and relative positional encoding (as proposed in [9]) – followed by a set of conformer [10] blocks which varies according to the dataset used for the downstream task (6 blocks for LRW, 12 blocks for LRS2). The conformer blocks feature the following parameters: $d^{\text{ff}} = 2048$, $n^{\text{head}} = 4$, $d^{\text{q}} = 256$, $d^{\text{k}} = 256$, $d^{\text{v}} = 256$; where d^{ff} is the hidden dimension of the feed-forward modules, n^{head} is the number of self-attention heads, and d^{q}, d^{k}, d^{v} are the dimensions of the key (K), query (Q), and value (V) in the self-attention layers respectively. The MLP consists of a linear layer with a hidden dimension of 256 units, ReLU activation, dropout, and a linear layer to project the representation to 256-dimensional latent space. For prediction, we average the PASE+ features, which are computed at 100 frames per second (fps), over time to match the frame rate of the input visual features (25 fps). We optimise our model using Adam ($\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$) combined with the Noam scheduler [39] (25 000 warm-up steps). The model is trained on LRS3 with a batch size of 32. For simplicity, we randomly sample 1 second from each clip and use it as the input to our network, discarding any utterances with less than 1 second in length.

3.5. Training settings in downstream tasks

LiRA-Supervised In LiRA-Supervised, we train word-level (Fig. 2a) and sentence-level lip-reading models (Fig. 2d) from scratch. In particular, for the task of word-level lip-reading, we add a MS-TCN followed by a linear classifier with an output dimension of 500 on top of the encoder like [17]. A cross-entropy loss is employed to optimise the whole model using Adam with decoupled Weight decay (AdamW) [14] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and a weight decay of 0.01 for 80 epochs with a batch size of 32. The initial learning rate is set to 0.0003. For the task of sentence-level lip-reading, we use 12 multi-head attention blocks ($d^{\text{ff}} = 2048$, $n^{\text{head}} = 4$, $d^{\text{q}} = 256$, $d^{\text{k}} = 256$, $d^{\text{v}} = 256$) together with a linear layer on the top of conformer blocks like [16]. Following [20], we use a combination of CTC and cross-entropy loss to train a hybrid CTC/Attention architecture for 50 epochs with a batch size of 8. In this case, we use Adam with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$ with the first 25 000 steps for warm-up. The initial learning rate is set to 0.0004. At the decoding phase, we use a beam size of 20 for beam search. During decoding, we also apply a transformer-based language model trained on LRS2, LRS3, and Librispeech 960h [25] (16.2 million words in total). Due to graphic memory limitations, we exclude utterances with more than 600 frames during training.

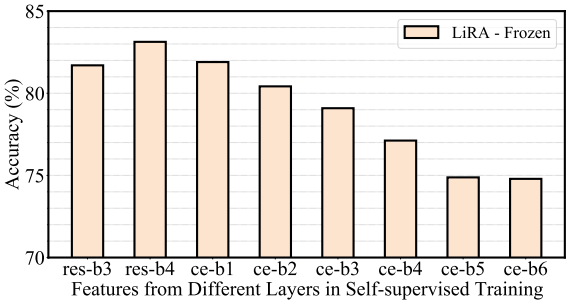


Figure 3: *Accuracy of feature classification (LiRA-Frozen) on LRW based on features extracted from different layers after pre-training on LRS3 via self-supervision. "res-b3" and "res-b4" refer to the output of blocks 3 and 4 from the ResNet-18 respectively; and "ce-b2" to "ce-b12" refer to the layers from every two conformer blocks from bottom to top.*

LiRA-Frozen At the end of self-supervised training, the features extracted from the pre-trained frozen encoder are fed to a classifier for evaluation. For word-level lip-reading, we use a MS-TCN, followed by a linear layer with an output size of 500 for classification (Fig. 2b). For the sentence-level lip-reading, the LiRA features are first fed to 12 conformer blocks, and then the encoded representations are used for CTC/attention joint training (Fig. 2e).

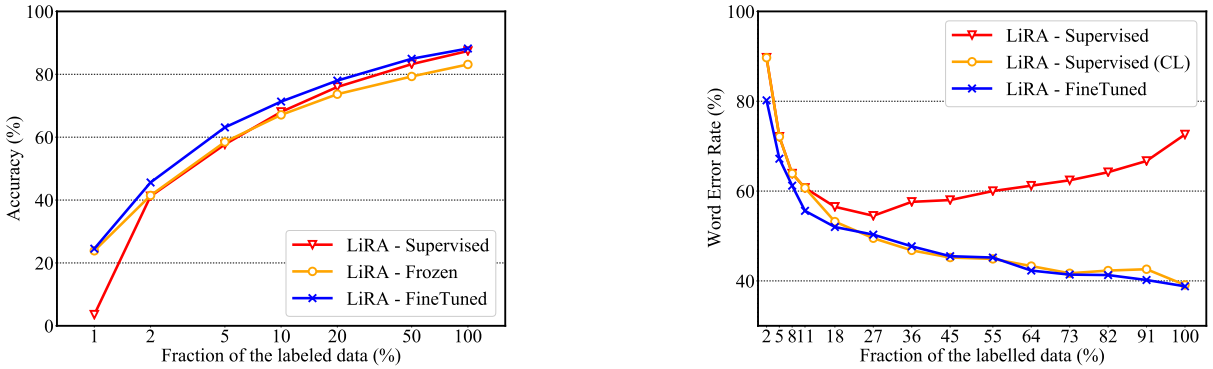
LiRA-FineTuned We follow the same hyperparameter setting as LiRA-Supervised, but instead of training from scratch, we initialise the encoder with the pre-trained weights from the pretext task and then fine-tune the entire model for word-level lip-reading (Fig. 2c) and sentence-level lip-reading (Fig. 2f).

4. Results

4.1. Word-level lip-reading

We first evaluate the performance of LiRA-Supervised by training the model from scratch. This leads to an accuracy of 87.4% on LRW which is very close to the state-of-art performance. For LiRA-Frozen, which is pre-trained on LRS3, the learnt visual speech representations are evaluated on word-level lip-reading by training a MS-TCN classifier on top of the frozen representations, as illustrated in Fig. 2b. Feature extraction performance (LiRA-Frozen) for different layers is portrayed in Fig. 3. We observe that the representations extracted from the last layer of the ResNet-18 achieve a maximum accuracy of 83.1 % as seen in Table 1. It is clear that the performance generally decreases as the layer becomes deeper, which may indicate that the features extracted in deeper layers are further tuned towards the pretext task and therefore fail to generalise as well for other tasks.

The performance of the 3 downstream scenarios while varying the amount of training data on LRW is shown in Fig. 4a. We use LRS3 for self-supervised pre-training. We observe that the feature extraction approach leads to superior performance compared to LiRA-Supervised when using smaller fractions of the labelled training set (1-2 %). This indicates that the pre-trained model learns useful visual features which also work well on LRW. By adopting this methodology, we can simply train the classification layers while the encoder remains frozen, and hence



(a) (b)

Figure 4: *Effect of the size of training data on downstream task performance. (a): Accuracy of the end-to-end model as a function of the percentage of the training set (on a logarithmic scale) used for training on LRW. (b) WER achieved by the end-to-end model as a function of the percentage of labelled data used for training on LRS2. All LiRA-Frozen and LiRA-FineTuned models are pre-trained on LRS3 via self-supervision. LiRA-Frozen models are trained using features extracted from the last layer of the ResNet-18 in the pre-trained model, since it achieves the best performance as demonstrated in Fig. 3. "CL" refers to the model being trained using curriculum learning. LRW and LRS2 contain 165 and 222 hours of labelled training data respectively.*

Table 2: *A comparison of the Word Error Rate (WER) between the baseline methods and ours (pre-trained on LRS3) on the LRS2 dataset. CL: Curriculum learning.*

Methods	Strategy	WER. (%)
Hyb. CTC/Att. [29]	Supervised	63.5
Conv-seq2seq [43]	Supervised	51.7
TDNN [41]	Supervised	48.9
TM-seq2seq [1]	Supervised	48.3
KD-seq2seq [3]	Unsupervised	51.3
LiRA-Supervised [16]	Supervised (CL)	39.1
LiRA-FineTuned	Self-supervised	38.8

significantly reduce the training time of our model. If we fine-tune the full model, including the encoder, then the performance improves further as shown in Fig. 4a.

We also observe that the gap between the performance of LiRA-FineTuned and LiRA-Supervised becomes smaller when we increase the amount of labelled data for training. This demonstrates that pre-training using the proposed self-supervised task is particularly beneficial when the labelled training set is very small. In the extreme case, where only 1 % of the labelled training data is used, LiRA-Supervised achieves an accuracy of 3.6 %. In contrast, we obtain 24.5 % accuracy when LiRA-FineTuned is trained using the same amount of data. This is mainly due to the fact that the self-supervised training provides a good initialisation for network training. We also show that LiRA-FineTuned provides an absolute improvement of 0.8 % in accuracy over LiRA-Supervised when both are trained on full LRW. This demonstrates that LiRA-FineTuned consistently outperforms LiRA-Supervised, even for larger labelled training sets.

4.2. Sentence-level lip-reading

To investigate the performance of visual speech representations in a more challenging task, we run training from scratch (Fig. 2d) and fine-tuning (Fig. 2f) experiments on LRS2 after pre-training on LRS3. We present our results as a function of the fraction of labelled data used during training.

Results are shown in Fig. 4b. It is evident that the performance of LiRA-FineTuned significantly outperforms the supervised baseline. We also observe that the performance of LiRA-Supervised is hard to optimise without a good initialisation. The performance becomes worse and worse as the training set increases beyond 18 % of the total amount of labelled data. This is likely due to the large amount of very long utterances featured in LRS2, which makes training from scratch especially difficult. To overcome this problem, we use curriculum learning. In particular, we first train the model using 11 % of the labelled training set, which is composed of videos with less than 155 frames in length and then use this model for initialisation when training on the entire training set. This curriculum learning strategy results in a substantially more effective training procedure, achieving 39.1 % WER for the full dataset.

Fine-tuning the self-supervised model leads to a small improvement over the curriculum learning strategy resulting in a 38.8 % WER. This is the new state-of-the-art performance on the LRS2 dataset when no external labelled datasets are used for training. We also observe that it leads to a 9.5 % absolute improvement compared to the previous state-of-the-art model [1], as reported in Table 2. Furthermore, as displayed in Fig. 4, we are able to outperform the previous state-of-the-art of 48.3 % WER using 18× fewer labelled data – 76 hours (36 % of LRS2) vs 1 362 hours (MVLRS, LRS2, and LRS3).

5. Conclusion

We present LiRA, which learns visual speech representations by cross-modal self-supervised learning. We train a visual model by predicting acoustic features from visual speech, and observe that it can be adapted for lip-reading with remarkable success. By fine-tuning our models for this new task, we achieve an accuracy of 88.1 % on LRW and report a WER of 38.8 % on LRS2. Given the extent of modern audiovisual corpora, we believe it would be promising to leverage this method towards other visual tasks such as emotion recognition and speaker recognition in the future.

6. Acknowledgements

The work of Pingchuan Ma has been partially supported by Honda. The work of Rodrigo Mira has been funded by Samsung. All datasets used in the experiments and all training, testing and ablation studies have been conducted at Imperial College.

7. References

[1] T. Afouras, J. S. Chung, *et al.*, "Deep audio-visual speech recognition," *IEEE PAMI*, 2018.

[2] T. Afouras, J. S. Chung, *et al.*, "LRS3-TED: A large-scale dataset for visual speech recognition," in *arXiv preprint arXiv:1809.00496*, 2018.

[3] T. Afouras, J. S. Chung, *et al.*, "ASR is all you need: Cross-modal distillation for lip reading," in *ICASSP*, 2020, pp. 2143–2147.

[4] H. Alwassel, D. Mahajan, *et al.*, "Self-supervised learning by cross-modal audio-video clustering," *CoRR*, vol. abs/1911.12667, 2019.

[5] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *ICCV*, 2017, pp. 609–617.

[6] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *ACCV*, vol. 10112, 2016, pp. 87–103.

[7] S. Chung, J. S. Chung, *et al.*, "Perfect match: Self-supervised embeddings for cross-modal retrieval," *J. Sel. Top. Signal Process.*, vol. 14, no. 3, pp. 568–576, 2020.

[8] S. Chung, H. Kang, *et al.*, "Seeing voices and hearing voices: Learning discriminative embeddings using cross-modal self-supervision," in *Interspeech*, 2020, pp. 3486–3490.

[9] Z. Dai, Z. Yang, *et al.*, "Transformer-XL: Attentive language models beyond a fixed-length context," in *ACL*, 2019, pp. 2978–2988.

[10] A. Gulati, J. Qin, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," pp. 5036–5040, 2020.

[11] D. E. King, "Dlib-ml: A machine learning toolkit," *JMLR*, vol. 10, pp. 1755–1758, 2009.

[12] A. Kolesnikov, X. Zhai, *et al.*, "Revisiting self-supervised visual representation learning," in *CVPR*, 2019, pp. 1920–1929.

[13] B. Korbar, D. Tran, *et al.*, "Cooperative learning of audio and video models from self-supervised synchronization," in *NeurIPS*, 2018, pp. 7774–7785.

[14] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.

[15] P. Ma, B. Martinez, *et al.*, "Towards practical lipreading with distilled and efficient models," in *ICASSP*, 2021, pp. 7608–7612.

[16] P. Ma, S. Petridis, *et al.*, "End-to-end audio-visual speech recognition with conformers," in *ICASSP*, 2021, pp. 7613–7617.

[17] P. Ma, Y. Wang, *et al.*, "Lip-reading with densely connected temporal convolutional networks," in *WACV*, 2021, pp. 2857–2866.

[18] T. Makino, H. Liao, *et al.*, "Recurrent neural network transducer for audio-visual speech recognition," in *ASRU*, 2019, pp. 905–912.

[19] B. Martínez, P. Ma, *et al.*, "Lipreading using temporal convolutional networks," in *ICASSP*, 2020, pp. 6319–6323.

[20] H. Miao, G. Cheng, *et al.*, "Online hybrid ctc/attention architecture for end-to-end speech recognition," in *Interspeech*, 2019, pp. 2623–2627.

[21] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *ECCV*, vol. 9910, 2016, pp. 69–84.

[22] A. van den Oord, Y. Li, *et al.*, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.

[23] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *ECCV*, vol. 11210, 2018, pp. 639–658.

[24] A. Owens, J. Wu, *et al.*, "Learning sight from sound: Ambient sound provides supervision for visual learning," *International Journal of Computer Vision*, vol. 126, no. 10, pp. 1120–1137, 2018.

[25] V. Panayotov, G. Chen, *et al.*, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.

[26] S. Pascual, M. Ravanelli, *et al.*, "Learning problem-agnostic speech representations from multiple self-supervised tasks," in *Interspeech*, 2019, pp. 161–165.

[27] S. Petridis, M. Pantic, *et al.*, "Prediction-based classification for audiovisual discrimination between laughter and speech," in *IEEE FG*, 2011, pp. 619–626.

[28] S. Petridis and M. Pantic, "Prediction-based audiovisual fusion for classification of non-linguistic vocalisations," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 45–58, 2016.

[29] S. Petridis, T. Stafylakis, *et al.*, "Audio-visual speech recognition with a hybrid ctc/attention architecture," in *SLT*, 2018, pp. 513–520.

[30] H. Pham, P. P. Liang, *et al.*, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *AAAI*, 2019, pp. 6892–6899.

[31] A. J. Piergiovanni, A. Angelova, *et al.*, "Evolving losses for unsupervised video representation learning," in *CVPR*, 2020, pp. 130–139.

[32] M. Ravanelli and Y. Bengio, "Learning speaker representations with mutual information," in *Interspeech*, 2019, pp. 1153–1157.

[33] M. Ravanelli, J. Zhong, *et al.*, "Multi-task self-supervised learning for robust speech recognition," in *ICASSP*, 2020, pp. 6989–6993.

[34] S. Schneider, A. Baevski, *et al.*, "Wav2vec: Unsupervised pre-training for speech recognition," in *Interspeech*, 2019, pp. 3465–3469.

[35] A. Shukla, S. Petridis, *et al.*, "Does visual self-supervision improve learning of speech representations?" *CoRR*, vol. abs/2005.01400, 2020.

[36] T. Stafylakis, M. H. Khan, *et al.*, "Pushing the boundaries of audiovisual word recognition using residual networks and LSTMs," *Computer Vision and Image Understanding*, vol. 176, pp. 22–32, 2018.

[37] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," in *Interspeech*, 2017, pp. 3652–3656.

[38] M. K. Tellamekala, M. F. Valstar, *et al.*, "Audio-visual predictive coding for self-supervised visual representation learning," in *ICPR*, 2020, pp. 9912–9919.

[39] A. Vaswani, N. Shazeer, *et al.*, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[40] X. Weng and K. Kitani, "Learning spatio-temporal features with two-stream deep 3D CNNs for lipreading," in *BMVC*, 2019, p. 269.

[41] J. Yu, S. Zhang, *et al.*, "Audio-visual recognition of overlapped speech for the LRS2 dataset," in *ICASSP*, 2020, pp. 6984–6988.

[42] R. Zhang, P. Isola, *et al.*, "Colorful image colorization," in *ECCV*, vol. 9907, 2016, pp. 649–666.

[43] X. Zhang, F. Cheng, *et al.*, "Spatio-temporal fusion based convolutional sequence learning for lip reading," in *ICCV*, 2019, pp. 713–722.