Cross-modal Self-Supervised Learning for Lip Reading: When Contrastive Learning meets Adversarial Training

Changchong Sheng
University of Oulu
Oulu, Finland
Changchong.Sheng@oulu.fi

Matti Pietikäinen
University of Oulu
Oulu, Finland
Matti.Pietikainen@oulu.fi

Qi Tian
Xidian University
Xi'an, China
wywqtian@gmail.com

Li Liu*
University of Oulu
Oulu, Finland
Li.Liu@oulu.fi

ABSTRACT

The goal of this work is to learn discriminative visual representations for lip reading without access to manual text annotation. Recent advances in cross-modal self-supervised learning have shown that the corresponding audio can serve as a supervisory signal to learn effective visual representations for lip reading. However, existing methods only exploit the natural synchronization of the video and the corresponding audio. We find that both video and audio are actually composed of speech-related information, identity-r elated information, and modal information. To make the visual representations (i) more discriminative for lip reading and (ii) indiscriminate with respect to the identities and modals, we propose a novel self-supervised learning framework called Adversarial Dual-Contrast Self-Supervised Learning (ADC-SSL), to go beyond previous methods by explicitly forcing the visual representations disentangled from speech-unrelated information. Experimental results clearly show that the proposed method outperforms state-of-the-art cross-modal self-supervised baselines by a large margin. Besides, ADC-SSL can outperform its supervised counterpart without any finetune.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; *Speech recognition*; **Neural networks**.

KEYWORDS

lip reading; cross-modal; self-supervised learning; adversarial training

*Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China.

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475415

ACM Reference Format:
Changchong Sheng, Matti Pietikäinen, Qi Tian, and Li Liu. 2021. Cross-modal Self-Supervised Learning for Lip Reading: When Contrastive Learning meets Adversarial Training. In *Proceedings of the 29th ACM Int'l Conference on Multimedia (MM '21), Oct. 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475415

1 INTRODUCTION

Supervised deep learning has brought revolutionary progress in many fields, such as image classification [17], object detection and segmentation [29], speech recognition [21], and machine translation [5]. Despite the remarkable progress witnessed in the past decade, the successes of supervised deep learning rely heavily on vast manually annotated training data, which has serious limitations in many real world applications including the interested lip reading task [1, 12] of this paper. Firstly, supervised learning is restricted to relatively narrow domains defined largely by the labeled training data, and thus leads to limited generalization ability. Secondly, a large amount of accurately labeled data like a large scale annotated dataset for lip reading is costly to gather, even extremely expensive for many applications like medical image analysis. Finally, However, for some specific tasks, *e.g.*, lip reading [1, 12], the cost of annotation can be extremely expensive. Recently, self-supervised learning has received a growing amount of attention due to its high label efficiency and good generalization. Self-supervised learning methods have shown great promise in natural language processing (*e.g.*, GPT [35, 36] and BERT [18]), computer vision (*e.g.*, CPC[26, 31], MOCO [24, 33], SimCLR [9, 10], RoCL [27] *et al.*) and cross-modal representation learning [3, 13, 15]. However, Methods that do not rely on massive accurate manual annotation like self-supervised learning are yet underexplored for the lip reading task.

How do we communicate with others? Literature in cognitive sciences demonstrates that humans rely both on hearing voices and seeing lip movements in the process of speech perception [6, 30]. We will be confused if the sound we hear does not match the lip movements we see, which means that voices and lip movements convey the same speech information. Motivated by this observation, voices and lip movements can naturally be treated as mutual supervisory signals to learn discriminative A-V representations for multiple downstream tasks, *e.g.*, cross-modal retrieval, speech recognition, and lip reading. In this work, we try to extract discriminative visual

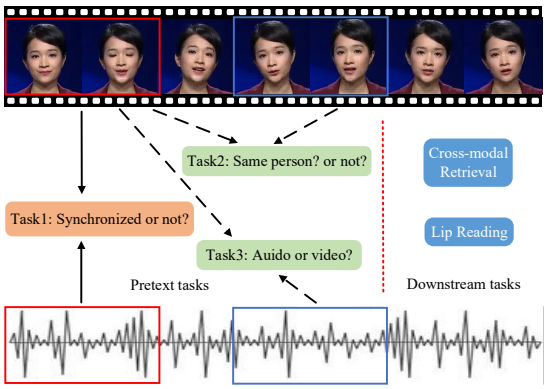

Figure 1: Problem & Framework description. This work aims to learn visual representations for lip reading. To do so, three pretext tasks are introduced: contrastive learning based on A-V synchronization; Identity adversarial training to make the representations free of identity-related information; Modal adversarial training to make the representations disentangled from modal information. *Solid lines:* Contrastive learning. *Dash lines:* Adversarial training.

representations for lip reading through a novel audio-visual cross-modal self-supervised learning method.

Given a talking face video, the lip movements and the audio are naturally co-occurring and synchronized. Previous works in this area try to use the pairwise contrastive strategy to force the visual embeddings closer to the corresponding audio embeddings and further apart from the non-corresponding audio embeddings [3, 13, 28, 32, 37]. Despite the remarkable progress, those methods have the following shortcomings. Firstly, the pairwise contrastive learning requires manual selection of the negative samples, and the performance depends heavily on the effectiveness of the negative samples. Secondly, representations learning only relies on the synchronized audio-video data pairs. Other self-supervisory signals, *e.g.*, speaker-related information and modal information, can also be utilized to optimize the quality of the learned representations. However, those self-supervisory signals are generally ignored in previous works.

To address these drawbacks, we present the Adversarial Dual-Contrast Self-Supervised (ADC-SSL) Framework to learn efficient visual representations by combining contrastive learning [31] and adversarial training [19], as illustrated in Figure 1. There are three pretext tasks: dual-contrastive learning based on A-V synchronization, identity adversarial training, and modal adversarial training.

Instead of the pairwise contrastive strategy used in previous works, another contrastive loss based on Noise Contrastive Estimation (NCE) [22] is considered in this paper. Compared to the pairwise objectives, NCE loss enforces that an embedding is far from multiple negative samples, instead of only one negative sample. Besides, we apply contrastive learning both on short-time and long-time A-V representations. This dual-contrast method can further optimize representations learning by integrating multi-scale speech information.

For the adversarial training, visual representations extracted from a single video share a common identity; Otherwise, the identity information is different. The objective is to force the learned visual representations to be free of identity information and modal information. To do so, we propose an identity discriminator and a modal classifier for A-V representations. The former's function is to discriminate whether the input visual features share a common identity; The latter is to predict whether the input feature belongs to visual modal or audio modal. Then adversarial training is achieved by Gradient Reversal Layer (GRL) [19]. We find that the original GRL is hard to balance these different training objectives. To solve this problem, the Momentum Gradient Reversal Layer (M-GRL) is proposed in this paper. M-GRL can optimize the training process by automatically learn the optimally weighted hyper-parameter based on the momentum update mechanism.

The major contributions of this work are summarized as follows.

- We propose a novel cross-modal self-supervised learning framework called Adversarial Dual-Contrast Self-Supervised Learning (ADC-SSL), which goes beyond previous methods by combing contrastive learning and adversarial training on three pretext tasks.
- We propose the Momentum Gradient Reversal Layer (M-GRL) for adversarial training, which stabilizes the training process by automatically learn the optimally weighted hyper-parameter.
- Experiments on cross-modal retrieval and lip reading clearly show that the proposed method outperforms state-of-the-art cross-modal self-supervised methods and exceeds the supervised counterparts both on the word-level and the sentence-level lip reading tasks.

2 RELATED WORK

2.1 Deep Lip Reading

The works on deep lip reading mainly focus on the architecture design of these two sub-networks: visual front-end networks and sequence back-end networks.

As for the design of visual front-end networks, plenty of works utilize deep CNNs to perform visual feature extraction. For example, Stafylakis *et al.*[38] proposes a simple variation of ResNet (changing the first 2D convolution layer to 3D convolution layer). This model consists of a shallow 3D CNN and deep 2D CNN, and it achieves 83% recognition accuracy for word-level lip reading on LRW [12] dataset. Due to the considerable performance of the model, most lip reading models [1, 34, 42] adopt it as the backbone network for visual features extraction. This network architecture is also used as the visual encoder in this work.

There are two main lip reading tasks: word-level and sentence-level. The former is to classify isolated words from the input videos, usually trained with multi-classification cross-entropy loss. Stafylakis *et al.* have created the baseline word-level lip reading model with Temporal convolution network (TCN) and BiLSTM based back-end network [38]. The latter is to do sentence-level sequence prediction, both connectionist temporal classification loss (CTC) [20] and sequence-to-sequence loss [39] can be used to train the

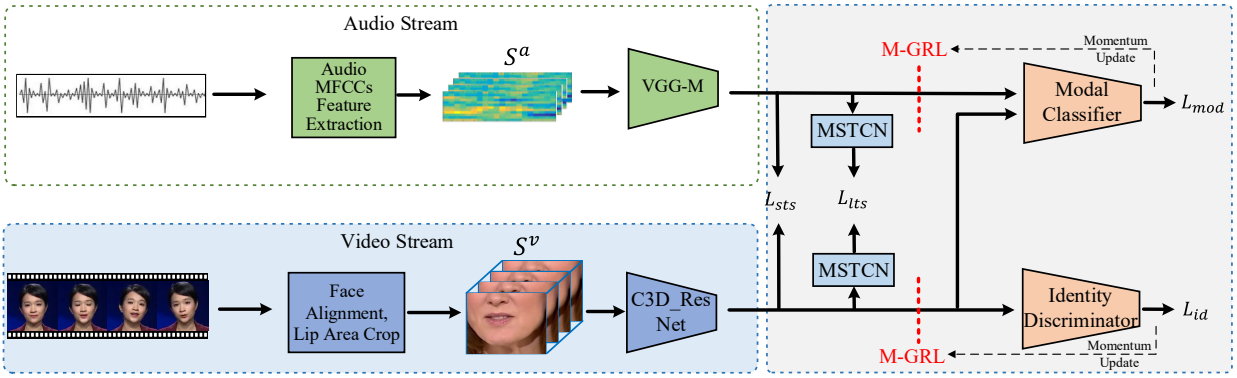

Figure 2: The overall pipeline of the proposed ADC-SSL framework. The left part are the pipelines of A-V features extraction, which are explained in Section 3.1. The right part is an illustration of the proposed three pretext tasks, which are explained in Section 3.2 and Section 3.3. *M-GRL:* Momentum Gradient Reverse Layer.

model. LipNet [4], consisting of 3D CNNs and BiGRUs, is the first end-to-end sentence-level lipreading model that simultaneously learns spatio-temporal visual features and temporal information. Besides, Afouras *et al.* introduce transformer self-attention architecture to lip reading. They propose Transformer-CTC model and Transformer-seq2seq model [1], and further discuss the difference between the two models in detail.

2.2 Audio-Visual Self-Supervised Learning

Audio-Visual self-supervised learning aims to extract efficient representations from the co-occurring A-V data pairs without any manual annotation. Based on the natural synchronization characteristics of audio and video, existing methods mainly adopt contrastive learning to achieve this goal. Chung *et al.*[13] is the first to train an A-V synchronization model in an end-to-end manner with margin-based [23] pairwise contrastive loss. Besides, they have also shown that the trained network works effectively for speaker detection and lip reading. With the same training strategy, Korbar *et al.*[28] broaden the scope of the study to encompass arbitrary human activities rather than lip movements. Except for margin-based loss, binary classification loss [3, 32, 37] is also widely used for A-V representations learning. Those works have proved the learned A-V representations can further transferred to more downstream tasks, such as visualizing the locations of sound sources, action recognition, audio-visual source separation, *et al.*. Recently, Chung *et al.*[15] reformulated the contrastive task as a multi-way matching task, and demonstrated that the use of multiple negative samples can improve the performance.

3 PROPOSED METHODOLOGY

In this section, we first give a brief introduction to the pipeline of the proposed ADC-SSL framework. And then, the framework is described in detail based on the three pretext tasks: A-V synchronization, identity adversarial training, and modal adversarial training. Finally, we elaborate on the network architectures used in this work.

3.1 The Overall Pipeline

As illustrated in Figure 2, given a talking mouth video S^v and its corresponding audio S^a, a visual encoder $f^v(\cdot)$ (C3D_ResNet) and an audio encoder $f^a(\cdot)$ (VGG-M) are first introduced to extract A-V embeddings. To ensure the consistency of A-V embeddings, both the audio encoder network and the visual encoder network ingest the clips with the same duration, generally 0.2 seconds [13, 15]. Specifically, the inputs to the audio encoder are the 13-dimensional Mel-frequency cepstral coefficients (MFCCs), extracted at every 10ms with 25ms frame length. And the input to the visual encoder is 5 consecutive mouth-centered cropped video ($fps = 25$) frames.

To learn effective visual representations for lip reading, three pretext tasks are introduced. Dual-Contrastive learning objectives L_{sts} and L_{lts} aim to make the visual embeddings closer to the corresponding audio embeddings both on short-time scale and long-time scale. Adversarial learning objectives L_{id} and L_{mod} make the learned embeddings indiscriminate for modal and identity information.

3.2 Dual-Contrastive Learning

As we mentioned above, most of the previous methods adopted a pairwise contrastive strategy to train the model, which suffers from hard negative mining. In addition, recent progress [9, 10, 15, 24] in self-supervised learning shows that the training significantly benefits from more negatives. Motivated by this, Noise Contrastive Estimation (NCE) [22] is considered as the training objective in this paper. NCE constructs a binary classification task, whose objective is to distinguish the real samples from the noise samples. In this paper, we build a contrastive loss based on NCE for the pretext task of A-V synchronization.

Let $\boldsymbol{h}^v_{1:T} = f^v(S^v)$ and $\boldsymbol{h}^a_{1:T} = f^a(S^a)$ denote visual representations and audio representations respectively, where T is time duration. And then, we randomly sample a minibatch of N examples and define the synchronization task on A-V pairs derived from the minibatch, resulting in $(2NT)$ A-V embeddings. Given a visual embedding $\boldsymbol{h}^v_{n,t}$ (as well as audio embedding) from a minibatch, we

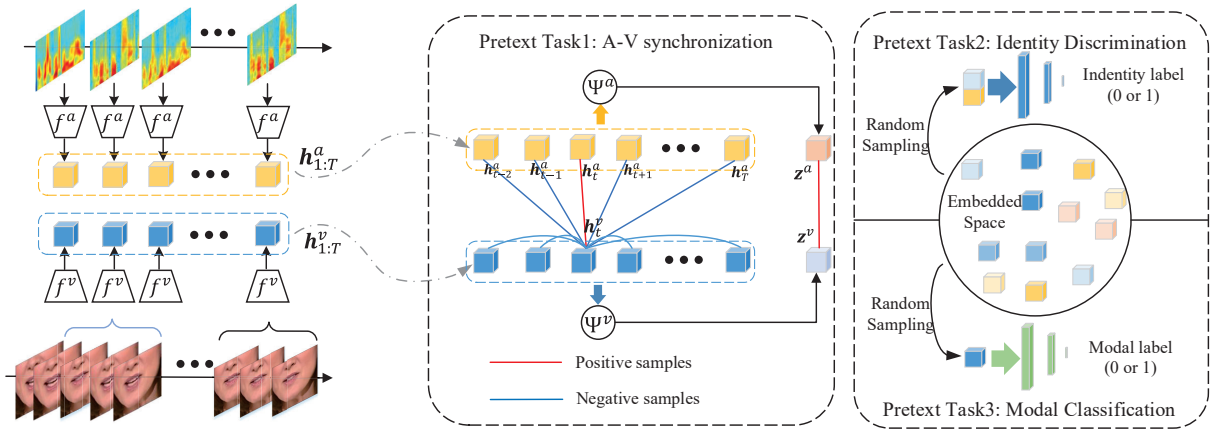

Figure 3: Those embeddings in embedded space denote the output of the visual encoder and the audio encoder. Pretext Task2: Identity discrimination. Pretext Task3: Modal classification.

treat the corresponding audio embedding $\boldsymbol{h}^a_{n,t}$ as positive sample, and the other $2(NT-1)$ A-V embeddings as negative samples. where n indicates the example index in a minibatch, and t denotes the timestep. We introduce $sim(\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{u}^T\boldsymbol{v}/\|\boldsymbol{u}\|\,\|\boldsymbol{v}\|$ to measure the similarity between two embeddings $\boldsymbol{u}$ and $\boldsymbol{v}$. Then the loss function on a positive pair $(\boldsymbol{h}^v_{n,t}, \boldsymbol{h}^a_{n,t})$ is defined as:

$$loss^{v,a}_{n,t} = -\log \frac{\exp(\text{sim}(\boldsymbol{h}^v_{n,t}, \boldsymbol{h}^a_{n,t})/\tau)}{\sum_{i,j}\exp(\text{sim}(\boldsymbol{h}^v_{n,t}, \boldsymbol{h}^{(a,v)}_{i,j})/\tau)} \tag{1}$$

where τ is a temperature hyper-parameter per [41]. In essence, this is simply a $(2NT-1)$-way cross-entropy loss that distinguishes the positive pair out of all other negative pairs within a minibatch. The loss encourages the positive pair to have a higher similarity than any negative pairs. This loss is based on the short-time A-V synchronization, and the total loss for short-time-synchronization L_{sts} in a minibatch is:

$$L_{sts} = \sum_{n,t}(loss^{v,a}_{n,t} + loss^{a,v}_{n,t}) \tag{2}$$

Compared to the manual selection of negatives and the complex curriculum learning strategy used in previous work [28], L_{sts} integrates both hard negatives (embeddings from the same video and audio) and easy negatives (embeddings from the other videos within a minibatch), which significantly simplify the training.

Actually, L_{sts} is performed based on the assumption of precise synchronization. However, the problem of off-sync sometimes occurs in videos [13]. L_{sts} performed on the off-sync videos may hurt the performance. Motivated by this observation, we propose to do contrastive learning for the whole video based on the speech matching. To do so, we introduce two multi-scale temporal convolution networks (MSTCN) with average pooling to aggregate global speech information for A-V representations, denoted as $\Psi^a(\cdot)$ and $\Psi^v(\cdot)$ respectively. As shown in Figure 3, let $\boldsymbol{z}^a = \Psi^a(\boldsymbol{h}^a_{1:T})$ and $\boldsymbol{z}^v = \Psi^v(\boldsymbol{h}^v_{1:T})$. Similar to L_{sts}, the long-time-synchronization L_{lts} can be defined as:

$$loss^{v,a}_{n} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}^v_n, \boldsymbol{z}^a_n)/\tau)}{\sum_i \exp(\text{sim}(\boldsymbol{z}^v_n, \boldsymbol{z}^{(a,v)}_i)/\tau)} \tag{3}$$

$$L_{lts} = \sum_n (loss^{v,a}_n + loss^{a,v}_n) \tag{4}$$

Based on the dual-contrastive learning mentioned above, the harness caused by off-sync examples can be mitigated to a large extent.

3.3 Adversarial Training

For the speech-related A-V representations learning, Zhou *et al.*[43] have demonstrated that explicitly disentangled training can make the learned representations more general to multiple speech-related downstream tasks. They have proposed the Disentangled Audio-Visual System (DAVS) to learn disentangled A-V representations. However, DAVS is performed based on multiple supervised labels, such as word label and identity label.

In this paper, we propose two novel and simple pretext tasks to force the learned A-V representations disentangled from identity-related information and modal information. The two adversarial pretext tasks (as illustrated in Figure 3) are performed based on adversarial training in a self-supervised manner. Next, we will briefly introduce the two pretext tasks and then explain how adversarial training is performed.

Identity Discrimination. Identity discrimination is based on the evidence that representations from a single video share a common identity. We build an identity discriminator $f^{id}(\cdot)$ whose objective is to discriminate whether the two input embeddings share a common identity or not. The two input embeddings are random sampled from the outputs of the visual encoder. Then, the identity discrimination loss L_{id} can be defined as:

$$\begin{aligned} L_{id} = &\frac{1}{K}\sum_i y_i \log(f^{id}(\boldsymbol{h}_i, \boldsymbol{h}'_i)) + \\ &(1-y_i)\log(1-f^{id}(\boldsymbol{h}_i, \boldsymbol{h}'_i)) \end{aligned} \tag{5}$$

Actually, this is a simple binary cross entropy loss used for 2-way classification. Where K is the total sampling number, $(\boldsymbol{h}_i, \boldsymbol{h}_i^{'})$ is the i_{th} sampling pairs, $y_i \in \{0, 1\}$ is the identity label.

Modal Classification. Similar to identity discrimination mentioned above, we build a modal classifier $f^{mod}(\cdot)$ whose objective is to discriminate whether the input embeddings are extracted from audio encoder or not. Then the modal classification loss L_{mod} is:

$$\begin{aligned} L_{mod} = & \frac{1}{2NT} \sum\nolimits_{n,t} y^v_{n,t} \log(f^{mod}(\boldsymbol{h}^v_{n,t})) + \\ & (1 - y^a_{n,t}) \log(1 - f^{mod}(\boldsymbol{h}^a_{n,t})) \end{aligned} \tag{6}$$

Momentum Gradient Reversal Layer. To enforce the representations disentangled from identity-related information and modal information, we propose a novel application of the Gradient Reversal Layer (GRL), originally introduced in [19] to learn domain-agnostic features. The GRL acts as the identity function during the forward pass of the network. On the gradient backward pass stage, the GRL reverses the weighted gradients flowing back from the corresponding branch. Inspired by this, we add GRL layers on the top of identity discriminator $f^{id}(\cdot)$ and modal classifier $f^{mod}(\cdot)$. So, The GRL inverts the sign of the weighted gradient that is backpropagated to the encoder networks $f^v(\cdot)$ and $f^a(\cdot)$.

By this means, the target of adversarial training is to do minimax learning, which can be written as:

$$\min_{\theta_{id}, \theta_{mod}} \max_{\theta_a, \theta_v} (L_{id} + L_{mod}) \tag{7}$$

Where θ_{id}, θ_{mod}, θ_a and θ_v are the parameters of $f^{id}(\cdot)$, $f^{mod}(\cdot)$, $f^a(\cdot)$ and $f^v(\cdot)$ respectively. Specifically, the A-V encoder networks are learned to maximize the L_{id} and L_{mod}, while the modal classifier and identity discriminator try to minimize the losses.

The gradient updates of θ_v can be written as:

$$\theta_v \leftarrow \theta_v - \mu(\frac{\partial(L_{sts} + L_{lts})}{\partial \theta_v} - \lambda_1 \frac{\partial L_{id}}{\partial \theta_v} - \lambda_2 \frac{\partial L_{mod}}{\partial \theta_v}) \tag{8}$$

where μ is the learning rate. λ_1 and λ_2 are weighted hyper-parameters applied on the GRL. We find that the fixed λ_1 and λ_2 will cause the training to become unstable or even not converge. In order to achieve a better balance between contrastive learning and adversarial training, we propose M-GRL to momentum update the weighted hyper-parameters λ_1 and λ_2.

Take the modal classifier as an example. We argue that the weighted hyper-parameters should be dynamically adjusted based on the modal classifier's uncertainty. When the uncertainty is high, the network should focus more on contrastive objectives. Otherwise, more attention should be paid to adversarial training. Specifically, we quantify the uncertainty as $H(f^{mod}) = -\sum_{c=1}^{C} p_c \log p_c$. The maximal value of $H(f^{mod})$ is $\log C$. Where C is the number of classes ($C = 2$) and p_c is the probability of class c. To do so, we reformulate λ_2 as:

$$\lambda_2 = \lambda_{high}(1 - e^{H(f^{mod}) - \log C}) + \lambda_{low} \tag{9}$$

Where λ_{high} and λ_{low} are constrained hyper-parameters. In experiments, we set $\lambda_{high} = 0.5$ and $\lambda_{low} = 0.001$. To ensure the stability of training, we update $H(f^{mod})$ with a momentum mechanism.

$$H(f^{mod}) \leftarrow mH(f^{mod}) + (1 - m)H_n(f^{mod}) \tag{10}$$

Here m is a momentum coefficient, and $H_n(f^{mod})$ is the uncertainty of the current minibatch. In this way, λ_2 can be automatically optimized to the optimal value. M-GRL is applied to the identity discriminator in the same way.

Overall Loss. With the combination of the dual-contrastive loss, the modal classification loss, and the identity discriminator loss, the final loss function of the proposed ADC-SSL training framework can be written as:

$$L_{ADC} = L_{sts} + L_{lts} + L_{id} + L_{mod} \tag{11}$$

The network is trained end-to-end with Eq. 11.

3.4 Network Architectures

In this section, we elaborate on all network architectures used in this work in detail.

Visual Encoder. We adopt a simple variation of ResNet34 [25, 38], called C3D_ResNet34 in this paper, as the visual encoder network $f^v(\cdot)$. C3D_ResNet34 only expands the first convolutional kernels to be 3D ones with the temporal receptive field equals to 5, and removes the last fully-connected layer.

Audio Encoder. Similar to [13, 15, 43], the audio encoder network $f^a(\cdot)$ is based on the VGG-M [8] CNN model, but the filter sizes are modified for the audio MFCCs features.

MSTCN. The function of $\Psi^v(\cdot)$ and $\Psi^a(\cdot)$ is to aggregate multi-scale speech information from the short-time representations. They consist of three stacked Multi-Scale dilated TCN layers, a fully connected (FC) layer, and an average pooling layer.

Identity Discriminator & Modal Classifier. The Identity Discriminator $f^{id}(\cdot)$ is stacked of a convolution layer, two linear layers, and a softmax activation layer. The convolution layer is to aggregate the two input embeddings. And the modal classifier $f^{mod}(\cdot)$ is composed of two linear layers and a softmax activation layer.

4 EXPERIMENTS AND ANALYSIS

In this section, we first describe the datasets used to evaluate our methods and some technical details for the self-supervised training stage. And then, the results and analysis on three downstream tasks are elaborated in detail.

Dataset	Subset	#Utter.	Word inst.	#Vocab.
LRW	Trainval	514k	514k	500
	Test	25k	25k	500
LRS2	Pretrain	96k	2M	41k
	Main	48k	344k	20k
LRS3	Pretrain	132k	4.2M	52k
	Trainval	32k	358k	17k
	Test	1,452	11k	2,136

Table 1: Description of the datasets used for training and testing.

4.1 Datasets and Technical Details

LRW. The LRW dataset [12] is commonly used for word-level visual speech classification task. It consists of up to 1000 utterances of 500 different English words, spoken by hundreds of different speakers.

Approach	V-A Retrieval		A-V Retrieval	
	R@1	R@10	R@1	R@10
Baseline [43]	64.2	84.7	67.7	85.8
Ours ($L_{sts} + L_{lts}$)	93.1	99.1	93.2	99.1
Ours($L_{sts} + L_{lts} + L_{mod}$)	**95.3**	**99.6**	**95.2**	**99.6**
Ours($L_{sts} + L_{lts} + L_{id}$)	91.8	98.9	92.3	99.0
Ours($L_{sts} + L_{lts} + L_{mod} + L_{id}$)	93.6	99.4	93.7	99.4

Table 2: 1:25000 A-V retrieval results with different training objectives.

The length of each video lasts 1.16 seconds (29 frames), and one word is uttered in the middle of the video.

LRS2 & LRS3. Both of the two datasets [11, 14] are commonly used for sentence-level lip reading task, containing three sets: *pre-train*, *trainval* and *test*. All videos in LRS2 are selected from BBC program, and it contains over 2.3 million words with a vocabulary size of 41,000. LRS3 are selected from TED and TEDx videos, it contains over 4.2 million words and the vocabulary size is 51,000.

The statistics of the datasets used in this paper is given in Table 1.

Technical details. For all the datasets, we use a face-alignment detector [7] to detect 68 facial landmark points for each video frame. For the input of visual encoder, a mouth-centered video of size 112×112 pixels is cropped based on the detected landmark points. The video inputs are converted to grayscale and all frames are normalized with respect to the overall mean and variance of all videos. Similar to [9], we also add projection heads that maps representations to the embedded space where contrastive loss is applied. For the hyper-parameters, temperature hyper-parameter τ is set as 0.07 [24], momentum coefficient $m = 0.99$. Standard Adam algorithm is implemented to optimize the parameters of the whole network. The Adam weight decay is 0.0001 and the Adam momentum is 0.9. We use the same data augmentation technique as that in [1] for visual input, such as horizontal flipping and random shifts.

4.2 The Effectiveness of the M-GRL.

As we pointed out in Sec. 3.3, the original GRL results instability or even non-convergence for the training of our network.

Here, we further elaborate the effectiveness of M-GRL. In Eq. 8, suppose we apply the original GRL here, λ_1 and λ_2 are constants. In the experiment, we find it difficult to set suitable values for λ_1 and λ_2. Take λ_2 as an example. If $\lambda_2 = 0.1$, L_{mod} almost converges to zero. That is to say, the modal adversarial training does not work at all. If $\lambda_2 = 0.5$, the training focuses too much on adversarial objective, and does not converge after a few iterations. Compared to the original GRL, the proposed M-GRL achieves a better balance between contrastive learning and adversarial training.

Fig. 4 lists the λ_1 & λ_2 curves in the training process on the LRW dataset. After some iterations, both λ_1 & λ_2 will converge to the optimal values that make the contrastive learning and adversarial training balanced.

4.3 Cross-modal Retrieval

Cross-modal retrieval task is used to evaluate the similarity between the A-V representations. We adopt the same evaluation protocols

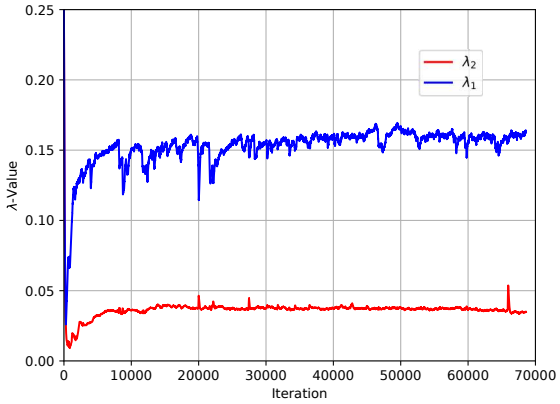

Figure 4: λ_1 & λ_2 curves in the first 70,000 iterations. *Red:* curves of λ_2, *blue:* curves of λ_1.

used in [43]. The cross-modal retrieval is performed on the test set of LRW (totally 25000 samples). Given a source video (audio), the objective is to find the matching audio (video) based on the cosine similarity of the representations. Here we report the R@1 and R@10 results. As we can see in Table 2, Our method significantly outperforms the baseline method.

Ablation study. To evaluate how the three pretext tasks impact the retrieval results, we also conduct several ablation experiments. We let the dual-contrastive learning as a baseline. Then ablation study is performed to evaluate the effects of identity adversarial training and modal adversarial training. As shown in Table 2, the best results are achieved based on dual-contrastive learning and modal adversarial training. However, the introduction of identity adversarial training has a side effect on cross-modal retrieval. This proves that in addition to speech information, identity information is also useful for cross-modal retrieval.

4.4 Word-level Lipreading

The goal of word-level lipreading on LRW is to recognize the isolated word class based on the input video. Experiments on this task are to show that the visual representations learned by the ADC-SSL framework are effective for lip reading. We compare the performance using the representations learned by the proposed self-supervised method to state-of-the-art self-supervised baselines, without any finetune on the visual encoder network. Besides, with

Training method	TOP-1 Acc (%)
SyncNet [13]	67.8
AVE-Net [3]	66.7
Perfect Match [15]	71.6
CDDL [16]	75.9
ADC-SSL (wo/ L_{sts})	71.4
ADC-SSL (wo/ L_{lts})	80.4
ADC-SSL (wo/ L_{mod})	82.7
ADC-SSL (wo/ L_{id})	82.9
ADC-SSL	83.9
ADC-SSL & finetune	**84.0**
Supervised Counterpart	79.1

Table 3: Word-level Lip reading Results. The Supervised Counterpart means the model (same as that used in the self-supervised training) trained from scratch.

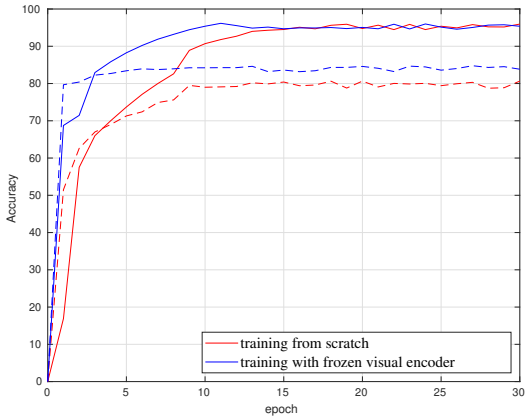

Figure 5: Training and validation accuracy curves. *Red:* curves of training from scratch, *blue:* curves of training with frozen visual encoder. *Solid:* curves of training accuracy, *Dash:* curves of validation accuracy.

the same network architecture, we also compare the performance with the full supervised counterpart trained from scratch.

The word-level lip reading network contains two sub-networks: a front-end visual encoder and a back-end sequence network. The front-end architecture is directly taken from the visual encoder $f^v(\cdot)$. For the back-end network, we propose a 2-layer temporal convolution network, followed by a 500-way softmax classification layer. This simple back-end classifier is widely used to evaluate the effectiveness of the learned visual representations [13, 15, 16]. We follow the common evaluation protocol, where only the back-end classifier is trained on top of the frozen visual encoder network, and test accuracy on LRW is used as a proxy for representation quality.

The results are listed in Table 3. Our ADC-SSL training method exceeds state-of-the-art self-supervised methods by a large margin. We also provide the results of ablation studies in this experiment. It turns out that all the four training objectives used in our framework are useful for the classification performance, where the short-time-synchronization L_{sts} contributes the most to the performance.

It is worth noting that our result (83.9%) even outperforms the supervised counterpart (79.1%). The training curves of these two methods are provided in Figure 5. As we can see, training accuracy is basically the same (about 95%) after 10 epochs. But the validation accuracy differs evidently (about 3%). This phenomenon suggests that compared to training from scratch, the representations learned by the self-supervised method can effectively prevent overfitting. Besides, we also list the result of fine-tuning the entire network based on the self-supervised pre-training. Its performance is not significantly improved compared to our results (83.9% *vs.* 84.0%).

4.5 Sentence-level Lipreading

Sentence-level lipreading aims to infer the content of a speech through the motion of the speaker's mouth. Compared to the word-level lipreading task, this task is more complicated and more practical. To further evaluate the quality of the learned visual representations, we provide the experiment results on this task. We would like to emphasize that there is currently no baseline method for self-supervised learning applied to this task, so we directly compare it with state-of-the-art end-to-end methods.

Next, the sequence back-end network is first presented. Then we give a brief introduction of the evaluation protocols and training details for this task.

Transformer back-end. For the sentence-level lip reading task, the output dimension is 39, including the 26 letters, 10 digitals, one punctuation "'" and [*SPACE*] and [*EOS*]. The commonly used transformer variant (transformer_seq2seq [1, 40]) network is adopted as the sequence back-end network. In this variant, we remove the embedding layer in the transformer encoder part because the input is visual representations instead of word class indexes. In addition, the output dimension of the last fully-connected layer of the decoder is changed to 39 to fit the size of the vocabulary.

Evaluation protocol. For all experiments, we report the Character Error Rate (CER) and Word Error Rate (WER). CER is defined as $CER = (S + D + I)/N$, where S, D and I are the numbers of substitutions, deletions, and insertions respectively to get from the reference to the hypothesis, and N is the number of characters in the reference. WER and CER are calculated in the same way. The difference lies in whether the formula is applied to character level or word level.

Training details. The pretrain sets of LRS2 and LRS3 are used to do self-supervised learning. After that, the parameters in the visual encoder are frozen. During the training on the transformer back-end, we follow a similar curriculum learning scheme as that in [11]: start training with utterances of 2 consequent words then gradually increase the number of words as training moving forward. Because the timestamp of every word in the input video is labeled, we can easily choose any continuous sentence instance within the dataset, and get the corresponding frames in the long input video. The model is first trained on the pretrain sets of LRS2 and LRS3 with text annotation. Then it is fine-tuned on the train-val sets of LRS2 and LRS3 separately.

Approach	Front-end architecture	Back-end architecture	LRS2		LRS3	
			CER	WER	CER	WER
WAS [11]	VGG-M	LSTM	-	70.4	-	-
TM-CTC [1]	C3D_ResNet34	Transformer	-	65.0	-	74.7
FC15 + CTC [1]	C3D_ResNet34	FC15	35.3	64.8	-	-
TM-seq2seq [1]	C3D_ResNet34	Transformer	38.6	**49.8**	-	59.9
CTC + KD [2]	Jasper-lip 5x3		-	-	-	60.9
Zhang *et al.*[42]	C3D_ResNet18	TF-block	-	51.7	-	60.1
Ours	C3D_ResNet34	Transformer	**35.1**	52.8	**40.5**	**59.2**
Supervised Counterpart	C3D_ResNet34	Transformer	41.4	60.2	48.1	68.8

Table 4: Sentence-level Lip reading Results (lower is better). *WER:* Word Error Rates. *CER:* Character Error Rates.

The transformer is trained with teacher forcing strategy. In the training process, the ground truth of the previous decoding step as the input to the decoder. During the inference stage, the decoder prediction at the last timestep is fed back to the decoder input. Decoding is performed with beam search of width 6. For a fair performance comparison, we do not use an external language model to optimize prediction results.

Comparative Evaluation. Results are presented in Table 4. Our ADC-SSL self-supervised method exceeds state-of-the-art fully supervised methods both on LRS2 dataset and LRS3 dataset, without any finetune on the visual encoder front-end.

It is worth noting that some of those SOTA methods (*e.g.*, TM-CTC, TM-seq2seq, CTC+KD, Zhang *et al.*[42]) need to pre-train on extra word-level lip reading datasets, *e.g.*, LRW dataset. Where TM-seq2seq and TM-CTC pre-train the visual front-end on the private word-level MV-LRS [14] dataset. Our proposed self-supervised training for this task in only performed on the pretrain dataset of LRS2 and LRS3. Besides, the results of CTC+KD [2] are achieved by distilling knowledge from an Automatic Speech Recognition (ASR) model that has been trained on a large-scale audio-only corpus.

To give a better comparison, we reproduce the supervised counterpart without extra datasets, and the results are listed on the penultimate column in Table 4. As we can see, Our self-supervised training method significantly outperforms that training from scratch. This also provides evidence for the conjecture that cross-modal self-supervised training can effectively prevent overfitting.

In sum, self-supervised training based on the ADC-SSL framework can extract effective visual representations for lip reading. Besides, the experiments and analysis of word-level lipreading task and sentence-level lipreading task proves that our proposed self-supervised training can effectively prevent overfitting.

5 CONCLUSIONS

In this paper, we proposed a new self-supervised training framework to learn discriminative visual representations for lip reading, without access to manual annotation. The proposed method combines contrastive learning and adversarial training by three pretext tasks, A-V synchronization, identity discrimination, and modal classification. In this way, the learned A-V representations are enforced to be free of identity-related information and modal-related information. Besides, a novel M-GRL is proposed to balance contrastive learning and adversarial training. Results on the cross-modal retrieval, word-level lipreading, and sentence-level lip reading tasks prove that the model trained with the proposed ADC-SSL framework outperforms state-of-the-art cross-modal self-supervised baselines, and even exceeds its supervised counterpart. The effectiveness of the ADC-SSL framework also opens up many possible applications for future works. For example, fake taking video detection, cross-modal anti-spoofing, lip movements generation *et al.*.

ACKNOWLEDGMENTS

This work was partially supported by the Academy of Finland under grant 331883, Outstanding Talents of "Ten Thousand Talents Plan" in Zhejiang Province (project no. 2018R51001), and the Natural Science Foundation of China (project no. 61976196). The authors also wish to acknowledge CSC IT Center for Science, Finland, for computational resources.

REFERENCES

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence* (2018).

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2020. 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)*. IEEE, 2143–2147.

[3] Relja Arandjelovic and Andrew Zisserman. 2018. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 435–451.

[4] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. 2016. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599* (2016).

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[6] Pascal Belin, Shirley Fecteau, and Catherine Bedard. 2004. Thinking the voice: neural correlates of voice perception. *Trends in cognitive sciences* 8, 3 (2004), 129–135.

[7] Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*. 1021–1030.

[8] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531* (2014).

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709* (2020).

[10] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. 2020. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029* (2020).

[11] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 3444–3453.

[12] Joon Son Chung and Andrew Zisserman. 2016. Lip reading in the wild. In *Asian Conference on Computer Vision*. Springer, 87–103.

[13] Joon Son Chung and Andrew Zisserman. 2016. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*. Springer, 251–263.

[14] J. S. Chung and A. Zisserman. 2017. Lip Reading in Profile. In *British Machine Vision Conference*.

[15] Soo-Whan Chung, Joon Son Chung, and Hong Goo Kang. 2020. Perfect Match: Self-Supervised Embeddings for Cross-modal Retrieval. *IEEE Journal of Selected Topics in Signal Processing* (2020).

[16] Soo-Whan Chung, Hong Goo Kang, and Joon Son Chung. 2020. Seeing voices and hearing voices: learning discriminative embeddings using cross-modal self-supervision. *arXiv preprint arXiv:2004.14326* (2020).

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[19] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17, 1 (2016), 2096–2030.

[20] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*. 369–376.

[21] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 6645–6649.

[22] Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 297–304.

[23] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. IEEE, 1735–1742.

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9729–9738.

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

[26] Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. 2019. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272* (2019).

[27] Minseon Kim, Jihoon Tack, and Sung Ju Hwang. 2020. Adversarial Self-Supervised Contrastive Learning. In *Thirty-fourth Conference on Neural Information Processing Systems, NeurIPS 2020*. NeurIPS.

[28] Bruno Korbar, Du Tran, and Lorenzo Torresani. 2018. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*. 7763–7774.

[29] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. 2020. Deep learning for generic object detection: A survey. *International journal of computer vision* 128, 2 (2020), 261–318.

[30] Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264, 5588 (1976), 746–748.

[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[32] Andrew Owens and Alexei A Efros. 2018. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 631–648.

[33] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. 2021. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11205–11214.

[34] Stavros Petridis, Themos Stafylakis, Pingehuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. 2018. End-to-end audiovisual speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6548–6552.

[35] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

[36] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[37] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. 2018. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4358–4366.

[38] Themos Stafylakis and Georgios Tzimiropoulos. 2017. Combining residual networks with LSTMs for lipreading. *arXiv preprint arXiv:1703.04105* (2017).

[39] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[41] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3733–3742.

[42] Xingxuan Zhang, Feng Cheng, and Shilin Wang. 2019. Spatio-temporal fusion based convolutional sequence learning for lip reading. In *Proceedings of the IEEE International Conference on Computer Vision*. 713–722.

[43] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9299–9306.