SynthVSR: Scaling Up Visual Speech Recognition With Synthetic Supervision

Xubo Liu[1*], Egor Lakomkin[2], Konstantinos Vougioukas[2], Pingchuan Ma[2], Honglie Chen[2], Ruiming Xie[2], Morrie Doulaty[2], Niko Moritz[2], Jachym Kolar[2], Stavros Petridis[2], Maja Pantic[2], Christian Fuegen[2]

[1]University of Surrey [2]Meta AI

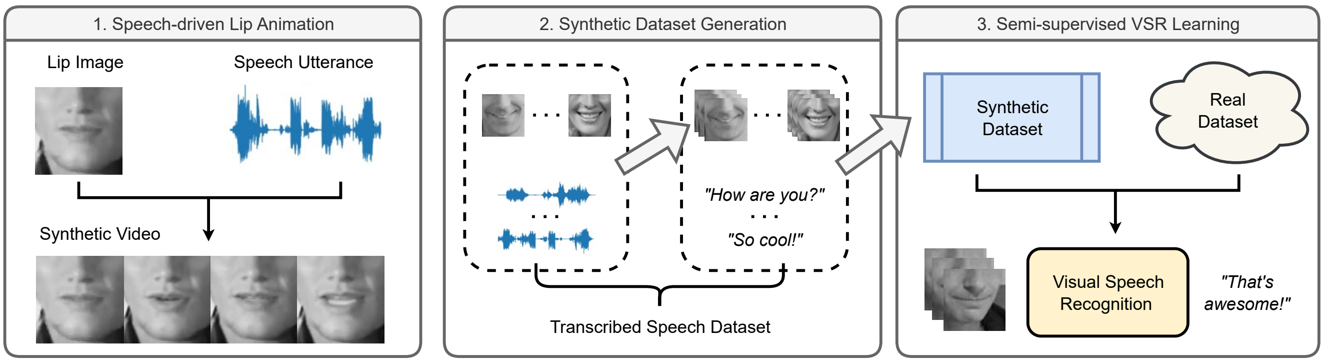

Figure 1. **Scaling up visual speech recognition with synthetic supervision (SynthVSR):** we propose SynthVSR, a semi-supervised framework that can substantially improve the performance of VSR models by using synthetic lip movements. Firstly, we introduce a speech-driven lip animation model that generates lip movement videos conditioned on input lip images and speech utterances (left). Secondly, we generate large-scale synthetic videos using transcribed speech datasets and lip images. The combination of synthetic videos and their corresponding speech transcriptions constitutes the synthetic dataset (centre). Finally, we conduct semi-supervised VSR training with synthetic and real datasets. Our method substantially improves the performance of VSR models with large-scale synthetic data (right).

Abstract

Recently reported state-of-the-art results in visual speech recognition (VSR) often rely on increasingly large amounts of video data, while the publicly available transcribed video datasets are limited in size. In this paper, for the first time, we study the potential of leveraging synthetic visual data for VSR. Our method, termed SynthVSR, substantially improves the performance of VSR systems with synthetic lip movements. The key idea behind SynthVSR is to leverage a speech-driven lip animation model that generates lip movements conditioned on the input speech. The speech-driven lip animation model is trained on an unlabeled audio-visual dataset and could be further optimized towards a pre-trained VSR model when labeled videos are available. As plenty of transcribed acoustic data and face images are available, we are able to generate large-scale synthetic data using the proposed lip animation model for semi-supervised VSR training. We evaluate the performance of our approach on the largest public VSR benchmark - Lip Reading Sentences 3 (LRS3). SynthVSR achieves a WER of 43.3% with only 30 hours of real labeled data, outperforming off-the-shelf approaches using thousands of hours of video. The WER is further reduced to 27.9% when using all 438 hours of labeled data from LRS3, which is on par with the state-of-the-art self-supervised AV-HuBERT method. Furthermore, when combined with large-scale pseudo-labeled audio-visual data SynthVSR yields a new state-of-the-art VSR WER of 16.9% using publicly available data only, surpassing the recent state-of-the-art approaches trained with 29 times more non-public machine-transcribed video data (90,000 hours). Finally, we perform extensive ablation studies to understand the effect of each component in our proposed method.

1. Introduction

Visual speech recognition (VSR), also known as lip reading, is the task of recognizing speech content based on visual lip movements. VSR has a wide range of applications in real-world scenarios such as helping the hearing-impaired perceive human speech and improving automatic speech recognition (ASR) in noisy environments.

VSR is a challenging task, as it requires capturing speech from high-dimensional spatio-temporal videos, while multiple words are visually ambiguous (e.g., "world" and

*Work done during an internship at Meta AI.

"word") in the visual streams. Recently, with the release of large-scale transcribed audio-visual datasets such as LRS2 [1] and LRS3 [2], deep neural networks have become the mainstream approach for VSR. However, even the largest public dataset for English VSR, LRS3, does not exceed 500 hours of transcribed video data. The lack of large-scale transcribed audio-visual datasets potentially results in VSR models which could only work in a laboratory environment i.e. limited word vocabulary and lip sources diversity [28].

A common solution to this issue is to collect and annotate large-scale audio-visual datasets. For example, [41,42] collected 90,000 hours of YouTube videos with user-uploaded transcriptions to achieve state-of-the-art performance on standard benchmarks. However, such a procedure is expensive and time-consuming, especially for most of the world's 7,000 languages [43]. If annotations are missing, the ASR can be used to generate the transcriptions automatically and this has been shown to be an effective approach to significantly improve VSR performance [28]. The other promising direction is to learn audio-visual speech representations from large amounts of parallel unlabeled audio-visual data in a self-supervised approach, and then fine-tune them on the limited labeled video dataset [43]. Nevertheless, publicly available video datasets are also limited and their usage may raise license-related[1] concerns, barring their use in commercial applications.

Human perception of speech is inherently multimodal, involving both audition and vision [43]. ASR, which is a complementary task to VSR, has achieved impressive performance in recent years, with tens of thousands of hours of annotated speech datasets [4, 20, 33] available for large-scale training. It is intuitive to ask: *Can we improve VSR with large amounts of transcribed acoustic-only ASR training data?* The key to this question is to take advantage of recent advances in speech-driven visual generative models [49,50]. By leveraging visual generative models, we can produce parallel synthetic videos for large-scale labeled audio datasets. Synthetic videos provide advantages such as having control over the target text and lip image as well as the duration of a generated utterance. To the best of our knowledge, the potential of leveraging synthetic visual data for improving VSR has never been studied in the literature.

In this work, we present SynthVSR, a novel semi-supervised framework for VSR. In particular, we first propose a speech-driven lip animation model that can generate synthetic lip movements video conditioned on the speech content. Next, the proposed lip animation model is used to generate synthetic video clips from transcribed speech datasets (e.g., Librispeech [33]) and human face datasets (e.g., CelebA [23]). Then, the synthetic videos together with the corresponding transcriptions are used in combination with the real video-text pairs (e.g., LRS3 [2]) for large-scale semi-supervised VSR training. The pipeline of SynthVSR is illustrated in Fig. 1. Unlike existing studies in exploiting unlabeled video data for VSR using methods such as pseudo-labeling [28] and self-supervised learning [43], we use the unlabeled audio-visual data to train a cross-modal generative model in order to bridge ASR training data and VSR. Furthermore, we propose to optimize the lip animation model towards a pre-trained VSR model when labeled videos are available. We empirically demonstrate that the semantically high level, spatio-temporal supervision signal from the pre-trained VSR model offers the lip animation model more accurate lip movements.

SynthVSR achieves remarkable performance gains with labeled video data of different scales. We evaluate the performance of SynthVSR on the largest public VSR benchmark LRS3 with a Conformer-Transformer encoder-decoder VSR model [28]. In the low-resource setup using only 30 hours of labeled video data from LRS3 [2], our approach achieves a VSR WER of 43.3%, substantially outperforming the former VSR methods using hundreds or thousands of hours of video data for supervised training [1,27,37,44,52] and self-supervised learning [3,26,43]. Notably, we demonstrate the first successful attempt that trains a VSR model with considerable WER performance using only 30 hours of real video data. Using the complete 438 hours from LRS3 further improves WER to 27.9% which is on par with the state-of-the-art self-supervised method AV-HuBERT-LARGE [43] that uses external 1,759 hours of unlabeled audio-visual data, but with fewer model parameters. Furthermore, following a recent high-resource setup [25] which uses additional 2,630 hours of ASR pseudo-labeled publicly available audio-visual data, our proposed method yields a new state-of-the-art VSR WER of 16.9%, surpassing the former state-of-the-art approaches [41, 42] trained on 90,000 hours of non-public machine-transcribed data.

Finally, we present extensive ablation studies to analyze where the improvement of SynthVSR comes from (e.g., diversity of lip sources, the scale of ASR data). We also show considerable VSR improvement using synthetic video data derived from Text-To-Speech (TTS)-generated speech, indicating the great potential of our method for VSR.

2. Related Work

Visual Speech Recognition. VSR has achieved remarkable progress with the success of deep learning and the availability of audio-visual datasets such as LRS2 [1] and LRS3 [2]. Progress was driven by adapting ASR approaches such as sequence-to-sequence models [11] and the Connectionist Temporal Classification (CTC) [5, 17] training objective. Recent studies have achieved noticeable results in VSR by using Transformer-based architecture [1], a convolutional variant [53], Conformer-based models [18], and

[1] Such as LRW [11] and LRS2 [1] datasets which are only permitted for the purpose of academic research.

an attention-based feature pooling method [35]. The recent advances in VSR models are mainly dependent on leveraging increasingly large-scale transcribed non-public video datasets such as LSVSR (3,800 hours) [44] and YT31k (31,000 hours) [28]. Some recent works [41,42] use 90,000 hours of non-publicly machine-transcribed video data to achieve state-of-the-art performance on standard benchmarks. Another popular trend is to leverage unlabeled audio-visual datasets such as AVSpeech [14] and VoxCeleb2 [10] with semi- or self-supervised learning. Using pre-trained ASR models to transcribe audio streams in unlabeled audio-visual datasets i.e., pseudo-labeling has demonstrated great potential to improve VSR performance. Such pseudo-labeled video data can be used for supervised training [28] or knowledge distillation [3]. Self-supervised approaches such as contrastive learning [12,29,32] and cross-modal feature prediction [26] have achieved significant improvement by exploiting unlabeled audio-visual data. AV-HuBERT [43] is the state-of-the-art self-supervised method that predicts iteratively refined cluster assignments from masked audio-visual streams and achieves impressive VSR performance, especially when only limited labeled video data is available (e.g., 30 hours).

Speech-Driven Facial Animation. Speech-driven facial animation aims to produce a talking face video conditioned on speech content. With the advances in deep visual generative models, speech-driven facial animation has significantly improved over the past years. Speech2Vid [9] is the first end-to-end neural network-based method for speech-driven facial animation. This method is optimized towards a pixel-level L_1 reconstruction loss which results in blurry generated frames. Recently, Generative Adversarial Networks (GANs) have made impressive progress [49,50] in facial animation, as they are able to generate sharp and highly textured facial frames. Initially, most of these methods synthesized frontal faces and focused on improving lip-sync [8, 50], preserving the identity of the target speaker [54]. As the generation of synthetic frontal faces improved, researchers began to focus on generating talking heads that could handle natural head movements [7, 55].

Learning from Synthetic Supervision. Training with synthetic data has attracted increasing research attention in recent years. Synthetic data could offer improved data diversity [13], which is useful in many tasks where labeled data is limited, such as image classification [16], objection detection [34, 36], semantic segmentation [40, 51], human pose estimation [13], and ASR [15, 31]. Unlike previous work, we focus on a new multimodal problem: how to produce and leverage synthetic video data for improving VSR.

3. SynthVSR

In this section, we introduce SynthVSR, a novel semi-supervised VSR framework to improve the performance of VSR models using synthetic data, as shown in Fig. 1. We propose a speech-driven lip animation model trained on an audio-visual dataset, which is used to produce synthetic videos with large-scale transcribed speech and face datasets for scaling up VSR training. We will introduce the baseline VSR model, the speech-driven lip animation model, and the semi-supervised VSR setting next.

3.1. Baseline VSR Model

The baseline VSR model we use in this work is based on [27, 28], which has achieved the state-of-the-art VSR performance on LRS3 [2] without the use of external data. The baseline VSR model is an encoder-decoder architecture. In particular, the encoder is comprised of two components, the visual front-end (a 3D convolutional layer followed by a ResNet-18 model [19,45]) and a Conformer [18] encoder. The decoder is based on the transformer architecture [47]. The baseline VSR model is trained end-to-end using a combination of the CTC loss [5,44] with an attention-based Cross-Entropy (CE) loss. Model details are described in the supplementary material.

3.2. Speech-driven Lip Animation

Inspired by the recent advances in speech-driven facial animation [49, 50, 54, 55], we propose an approach for speech-driven lip animation that generates videos of talking mouth regions conditioned on speech utterances. The output space of the lip animation model is the same as the VSR input space. The proposed lip animation model is based on a temporal GAN [49,50] with two discriminators. We further propose a VSR perceptual loss when labeled video data is available. The architecture of the speech-driven lip animation model is illustrated in the left part of Fig. 2. We will introduce each component in the next sections.

Generator. The generator G is an encoder-decoder structure, as shown in the right part of the Fig. 2. The generator uses the first frame of a video clip, a speech clip, and a head rotation sequence as inputs. The head rotation is used as the additional condition which helps better model the lip movements as most training videos are not static. The speech clip is divided into overlapping 200 ms chunks with a stride of 40 ms. The generator produces the corresponding video frame for each speech chunk. In the generator, an image encoder E_i and a speech encoder E_s are used to capture the visual information and speech context into latent embeddings z_i and z_s, respectively. The image encoder E_i uses a stack of 2D convolutional layers to extract the visual embedding. The speech encoder E_s comprises of a stack of 1D convolutional layers followed by a stack of GRU layers. The head rotations are provided in the form of sequences of 3D rotation matrices [55] $z_r \in \mathbb{R}^{3\times3}$ with respect to the first frame. The three embeddings z_i, z_s and z_r are concatenated and used to modulate the convolutional lay-

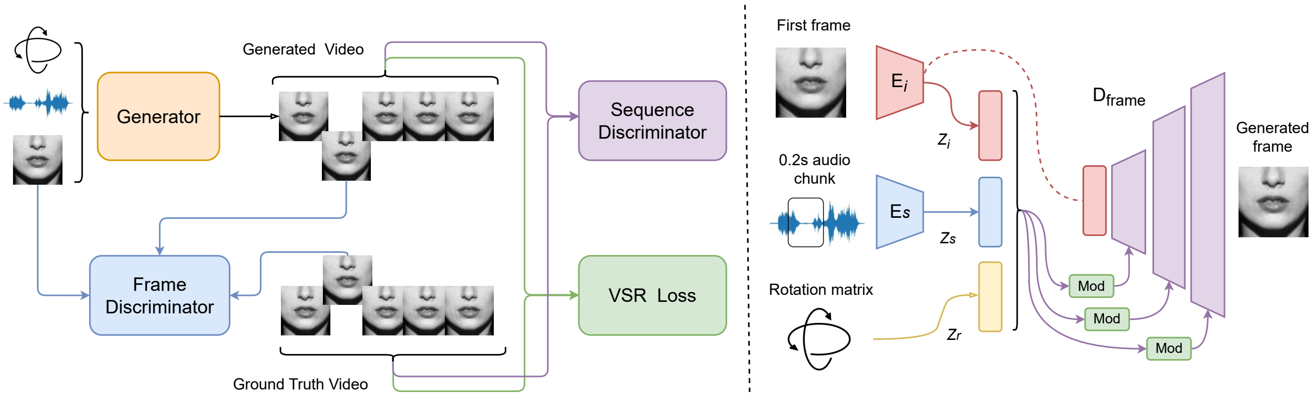

Figure 2. **Architecture of proposed speech-driven lip animation model.** Left: GAN-based speech-driven lip animation model generating lip movements given a lip image, a speech utterance, and a rotation sequence; Right: structure of the generator in the lip animation model.

ers in the frame decoder D_{frame}, which is similar to StyleGAN2 [48]. Different from StyleGAN2, which generates frames from a learned constant input, we use the penultimate layer features of the image encoder E_i as the input.

Discriminators. The speech-driven lip animation system has two discriminators: frame discriminator D_{img} and sequence discriminator D_{seq}. The frame discriminator is a stack of CNN layers that operates on the image frame level. The original first frame is concatenated channel-wise to the target frame to form the input of the frame discriminator. This helps enforce visual consistency. The sequence discriminator operates on the sequence level to ensure the temporal consistency of synthetic lip movements. The sequence discriminator uses spatio-temporal convolutions to encode the image sequence, followed by GRU layers to determine if the sequence is real or not. Specifically, the frame discriminator D_{img} is trained on frames that are uniformly sampled from a video v using a sampling function $S(v)$. The first frame v_1 is fed to the D_{img} as the condition. The input speech signal is s. The adversarial loss of the D_{img} is defined as:

$$\begin{aligned} \mathcal{L}_{Disc}^{img} &= \mathbb{E}_v[\log D_{img}(S(v), v_1)] \\ &+ \mathbb{E}_{v,s}[\log(1 - D_{img}(S(G(s, v_1)), v_1)]. \end{aligned} \tag{1}$$

D_{seq} operates on the entire sequence video v. The adversarial loss of the D_{seq} is defined as follows:

$$\begin{aligned} \mathcal{L}_{Disc}^{seq} &= \mathbb{E}_v[\log D_{seq}(v)] + \\ &\mathbb{E}_{v,s}[\log(1 - D_{seq}(G(s, v_1))]. \end{aligned} \tag{2}$$

VSR Perceptual Loss. We further propose to optimize the lip animation model towards a VSR perceptual loss if labeled video data is available. We first pre-train a VSR model as introduced in Sec. 3.1. The proposed VSR perceptual loss corresponds to a weighted sum of feature distances computed from the visual front-end and the Transformer decoder of the pre-trained VSR model for real and generated samples. We use L_1 norm to measure the visual embedding distance and Kullback–Leibler (KL) divergence to measure the logits distribution distance, respectively. The VSR perceptual loss is obtained by:

$$\mathcal{L}_{VSR} = \lambda_{visual} \left\| z_f^r - z_f^s \right\|_1 + \lambda_{logits} \, \mathrm{KL}(\hat{y}^r, \hat{y}^s), \tag{3}$$

where z_f^r and z_f^s are the VSR front-end visual features of the real and synthetic video, respectively, $\hat{y}^r$ and $\hat{y}^s$ is the VSR predicted logits distribution of real and synthetic video, respectively, λ_{visual} and λ_{logits} control the weights of these two perceptual losses. The VSR model is frozen during the lip animation model training.

Training Objectives. The speech-driven lip animation model is trained using a combination of a reconstruction loss, adversarial losses, and a VSR perceptual loss. The reconstruction loss is computed based on the L_1 distance between the generated video $\hat{v}$ and ground truth video v:

$$\mathcal{L}_{rec} = \left\| v - \hat{v} \right\|_1. \tag{4}$$

The overall training loss for the lip animation model is:

$$\begin{aligned} \mathcal{L}_{Animation} = \lambda_{disc}^{img} \mathcal{L}_{disc}^{img} + \lambda_{disc}^{seq} \mathcal{L}_{disc}^{seq} \\ + \lambda_{rec} \mathcal{L}_{rec} + \mathcal{L}_{VSR}, \end{aligned} \tag{5}$$

where λ_{disc}^{img}, λ_{disc}^{seq}, and λ_{rec} represent the coefficient of the adversarial loss of frame discriminator D_{img}, the adversarial loss of sequence discriminator D_{seq}, and the reconstruction loss, respectively.

3.3. Semi-supervised VSR with Synthetic Data

In a typical supervised setting, a VSR model is trained from the labeled dataset $\mathcal{D}_{real} = \{(v_i^r, y_i^r)\}_{i=1}^{n_r}$, where n_r is the number of paired video clips v_i^r and its transcriptions y_i^r. In SynthVSR, we first train a speech-driven lip animation model on an unlabeled audio-visual dataset $\mathcal{D}_{av} = \{(v_i^{av}, s_i^{av})\}_{i=1}^{n_{av}}$, where n_{av} is the number of paired video clips v_i^{av} and speech utterances s_i^{av}. Furthermore, we can use the trained generator G to generate synthetic video from a labeled speech dataset $\mathcal{D}_s = \{(s_i^s, y_i^s)\}_{i=1}^{n_s}$ and a cropped lips dataset $\mathcal{D}_f = \{c_i^f\}_{i=1}^{n_f}$, where n_s is the number of paired speech clips s_i^s and its transcriptions y_i^s and n_f is the number of lip images c_i^f. For each speech clip s_i^s, the generator G generates its parallel synthetic video $\hat{v}_i^s$:

$$G(s_i^s, S(\mathcal{D}_f)) \mapsto \hat{v}_i^s, \tag{6}$$

where $S(\cdot)$ is a uniform sampling function. Then, the synthetic dataset $\mathcal{D}_{synth} = \{\hat{v}_i^s, y_i^s)\}_{i=1}^{n_s}$ is obtained, where in general $n_s >> n_r$ as there are far more annotated speech datasets (e.g., Librispeech [33]) than labeled video datasets. Finally, the VSR model is trained from $\mathcal{D}_{real} \cup \mathcal{D}_{synth}$, where the rich visual and text label information in $\mathcal{D}_{synth}$ offers significant performance improvement for VSR.

4. Experiments

4.1. Datasets

We use several public datasets in this work. (1) Audio-visual datasets: LRS3 [2], AVSpeech [14] and VoxCeleb2 [10]; (2) Speech datasets: Librispeech [33], TED-LIUM 3 [20], Common Voice [4]; (3) Facial dataset: CelebA [23].

LRS3. We conduct experiments on the LRS3 [2] dataset, which is the largest public benchmark for English VSR containing 438.9 hours of video clips from TED talks (408, 30, and 0.9 hours in the pre-training, training-validation, and test set, respectively).

AVSpeech & VoxCeleb2. A recent work [25] uses public language classifier [46] and ASR models [6] to filter for English content and obtain pseudo-transcriptions for two multilingual audio-visual datasets: AVSpeech [14] and VoxCeleb2 [10], containing 1323 hours from AVSpeech and 1307 hours from VoxCeleb2. We follow the same data setting for our high-resource setting. Furthermore, we filter out AVSpeech videos with large jitter resulting in 933 hours of videos overall for speech-driven lip animation training.

Datasets for Speech-driven Lip Animation Training. The lip animation model is trained on a combination of LRS3 (pre-training and training-validation splits) and the English subset (933 hours) of AVSpeech datasets.

Datasets for Synthetic Data Generation. We use Librispeech [33], TED-LIUM 3 [20], Common Voice (English split) [4] datasets as the speech sources. We split Librispeech audio clips into segments of less than 6 seconds based on the provided silence annotations. For TED-LIUM 3 and Common Voice, we filter out the audio clips which are longer than 20s. As a result, 944 hours, 465 hours, and 2,243 hours of speech data are obtained for Librispeech, TED-LIUM 3, and Common Voice, respectively. We use the CelebA [23] dataset as a source of lip images, which has 202,599 face images and 10,177 identities. For each speech clip, we randomly sample one image from CelebA to generate one synthetic video. We use a static rotation matrix in the generation process. In total, 3652 hours of synthetic video clips are generated for scaling up VSR training.

4.2. Data Processing

We train a mouth detection module using the Faster R-CNN object detection architecture [38] on the CelebA [23] dataset. We leverage the mouth corner positions provided in the dataset to identify the area around the mouth in each video frame. With this network, we locate and extract a 96×96 bounding box around the mouth, which we subsequently convert to grayscale. We use the same video pre-processing methods for lip animation and VSR models. For the text vocabulary, we use SentencePiece [22] 5,000 subword units with a vocabulary size of 5,000.

4.3. Implementation Details

VSR Model. We consider two model configurations: (1) Conformer-BASE (250M) with 12-layer Conformer encoder, 6-layer Transformer decoder, 768 input dimensions, 3,072 feed-forward dimensions, and 16 attention heads; (2) Conformer-LARGE (783M) with 24-layers Conformer encoder, 9-layer Transformer decoder, 1024 input dimensions, 4,096 feed-forward dimensions, and 16 attention heads. For each configuration, the encoder and decoder have the same dimensions and attention heads.

We use horizontal flipping, random cropping, and adaptive time masking as the VSR data augmentation methods [28]. We train the BASE and LARGE models for 75 epochs on 64 A100-GPUs with the AdamW [24] optimizer, a cosine learning rate scheduler, and a warm-up of 5 epochs. The peak learning rate is 1×10^{-3} and 8×10^{-4} for the BASE and LARGE models respectively. The number of frames in each batch is limited to 2,400 and 1,600 frames for the BASE and LARGE models, respectively. Following [25, 28], a pre-trained transformer-based language model is used in the VSR decoding stage. The training configuration is the same as that used in the previous work [25].

Speech-driven Lip Animation. For each training video clip, we sample a sequence of 75 frames that corresponds to 3 seconds. The first frame of the sampled video sequence is used to drive the lip animation. We train the lip animation model for 70 epochs on 32 A100-GPUs with a batch size of 3 per GPU. The details of the module structure in the lip animation model are described in the supplementary material. Adam [21] optimizer is used to train the lip animation model with the learning rate 1×10^{-4} for the generator and the frame discriminator. Sequence discriminator uses a smaller learning rate of 1×10^{-5}. The weights for each loss term are $\lambda_{disc}^{img} = 1$, $\lambda_{disc}^{seq} = 0.2$, $\lambda_{rec} = 300$ in Eq. (5).

We consider five lip animation model configurations. The lip animation model without VSR perceptual loss is referred to LAM-Baseline. The lip animation model with the BASE VSR model trained on LRS3 is referred to as LAM-

Method	Backbone	LM	Labeled data (hrs)	Unlabeled data (hrs)	Synthetic data (hrs)	WER (%)
Afouras et al. [3]	CNN	✓	595‡	334	-	59.8
Ren et al. [37]	Transformer	✗	818‡	-	-	59.0
Afouras et al. [1]	Transformer	✓	1,519†‡	-	-	58.9
Xu et al. [52]	RNN	✗	595‡	-	-	57.8
Shillingford et al. [44]	RNN	✓	3,886†	-	-	55.1
Ma et al. [26]	Transformer	✗	433	1,759	-	49.6*
Ma et al. [27]	Conformer	✓	438	-	-	46.9
AV-HuBERT-BASE [43]	Transformer	✗	30	1,759	-	46.1
SynthVSR	Conformer-BASE	✗	30	-	-	104.0
		✗	-	-	3,652	100.3
		✗	30	-	3,652	44.7
		✓	30	-	3,652	**43.3**

Table 1. Experimental results of low-resource labeled data setting on LRS3 (test). LM denotes whether or not a language model is used in the decoding. †Includes non-publicly available data. ‡Includes datasets that are only permitted for the purpose of academic research. hrs is an abbreviation for hours. *Result taken from [43].

LRS3-VSR-VL, with the $\lambda_{visual} = 250$ and $\lambda_{logits} = 10$ in Eq. (3). Two variants LAM-LRS3-VSR-L and LAM-LRS3-VSR-V are further designed with the $\lambda_{visual} = 0$ and $\lambda_{logits} = 0$, respectively. These two model configurations are used in ablation studies. Last, the lip animation model with the BASE VSR model trained on LRS3 and 2,630 hours of pseudo-labeled AVSpeech and VoxCeleb2 is referred to as LAM-LRS3-AVoX-VSR, with the $\lambda_{visual} = 500$ and $\lambda_{logits} = 10$.

4.4. Low-resource Labeled Data Setting

The LAM-Baseline model is used to generate 3,652 hours of synthetic data. We use 30 hours of LRS3 (training-validation) to evaluate the VSR performance when labeled data is scarce. We conduct experiments on the BASE VSR model. The results are reported in Tab. 1. As supervised VSR training from scratch with long sequences often poses optimization problems [1, 11, 28], we first use 30 hours of LRS3 and the 944 hours of Librispeech synthetic data to pre-train a VSR model with the same architecture as [28], then we take the weights of the pre-trained visual front-end to train the BASE VSR model. The pre-training details are described in the supplementary material.

Training with synthetic data leads to dramatic performance improvement when only 30 hours of labeled data is available. Using 3,652 hours of synthetic data and 30 hours of LRS3 labeled data for training, the BASE model achieves the WER 43.3% (44.7% w./o. language model). Training from 30 hours of LRS3 results in a poor WER of 104%. Our proposed method significantly outperforms the former approaches using hundreds or thousands of data for supervised training [1,27,37,44,52] and self-supervised learning [3, 26]. Furthermore, our method performs better than the AV-HuBERT-BASE [43] model (46.1%) that uses 30 hours of LRS3 and 1,759 hours of unlabeled audio-visual data for self-supervised learning. Notably, we show the first attempt that trains a VSR model with considerable WER performance using only 30 hours of real video data. In addition, we observe that training only from synthetic data performs poorly (100.3%), which may be caused by the domain mismatch between the real and synthetic data.

4.5. LRS3 Labeled Data Setting

In this setting, we report the results when using the full 438 hours of LRS3. Experiments are conducted on the BASE VSR model. To avoid the optimization problem, we initialize the weights of the visual front-end from a publicly available Conformer-based model [28] pre-trained on LRS3 using curriculum learning. The results are shown in Tab. 2.

Our BASE model achieves WER 36.7% when using 438 hours of LRS3 data for training. We generate 3,652 hours of synthetic data using the LAM-LRS3-VSR-VL model. Using 3,652 hours of synthetic data and 438 hours of LRS3 labeled data, the BASE VSR model achieves the WER 27.9% (28.4% w/o language model, corresponding to a WER reduction of 8.3%). Our method outperforms three recent approaches using 31,000 (33.6%), 1,459 (31.5%), and 2,679 (30.7%) hours of labeled data, respectively. When compared with the state-of-the-art self-supervised method AV-HuBERT [43] that uses additional 1,759 hours of unlabeled audio-visual data, our method outperforms the AV-HuBERT-BASE model (34.8%) by a large margin. Our method slightly performs better than the AV-HuBERT-LARGE model (28.6%), but with fewer model parameters (our BASE model 250M vs AV-HuBERT-LARGE 390M). Note that we compare with the AV-HuBERT results without self-training as we do not use the pseudo-labeled 933 hours of AVSpeech subset for VSR training.

4.6. High-resource Labeled Data Setting

We further evaluate the scalability of SynthVSR: when using machine-transcribed AVSpeech [14] and VoxCeleb2

Method	Backbone	LM	Labeled data (hrs)	Unlabeled data (hrs)	Synthetic data (hrs)	WER (%)
AV-HuBERT-BASE [43]	Transformer	✗	433	1,759	-	34.8
Makino et al. [30]	Transformer	✗	31,000†	-	-	33.6
Ma et al. [28]	Conformer	✓	1,459‡	-	-	31.5
Prajwal et al. [35]	Transformer	✓	2,676†	-	-	30.7
AV-HuBERT-LARGE [43]	Transformer	✓	433	1,759	-	28.6
AV-HuBERT-LARGE w. Self-Training [43]	Transformer	✓	433	1,759	-	26.9
Auto-AVSR [25]	Conformer	✓	3,448‡	-	-	19.1
Serdyuk et al. [42]	Transformer	✗	90,000†	-	-	25.9
Serdyuk et al. [41]	Transformer	✗	90,000†	-	-	17.0
SynthVSR	Conformer-BASE	✗	438	-	-	36.7
		✗	438	-	3,652	28.4
		✓	438	-	3,652	**27.9**
		✗	3,068	-	-	21.2
		✗	3,068	-	3,652	19.4
		✓	3,068	-	3,652	**18.7**
SynthVSR	Conformer-LARGE	✗	3,068	-	3,652	18.2
		✓	3,068	-	3,652	**16.9**

Table 2. Experimental results of LRS3 & high-resource labeled data setting on LRS3 (test). LM denotes whether or not a language model is used in the decoding. †Includes non-publicly available data. ‡Includes datasets that are only permitted for the purpose of academic research. hrs is an abbreviation for hours.

[10] as additional training data. We generate 3,652 hours of synthetic data with LAM-LRS3-AVoX-VSR and conduct experiments on BASE and LARGE models. The visual front-end is initialized from a pre-trained model which is the same as that used in the LRS3 labeled data setting.

We first train the BASE model with 438 hours of labeled LRS3 and 2,630 hours of pseudo-labeled data, resulting in a strong VSR baseline with the WER 21.2%. By using additional 3,652 hours of synthetic data, the WER of the BASE model improves to 18.7% (19.4% w./o. language model), which outperforms [25] that uses additional labeled dataset LRS2 (223 hours) and LRW (157 hours) for training. Although the VSR model has seen a large amount of labeled data, and the speech-driven lip animation model is trained from part of the VSR training data, synthetic data can still lead to considerable performance gains. Furthermore, increasing the model size from BASE to LARGE results in better VSR performance with the WER of 16.9% (18.2% w./o. language model), which is the current state-of-the-art performance on LRS3, with publicly available data only.

4.7. Why SynthVSR Improves VSR?

We conduct extensive ablation studies to understand the impact of VSR perceptual loss, diversity of lip sources, and the scale of ASR data for SynthVSR. In addition, we assess the domain mismatch between the real and synthetic data using VSR models. We perform experiments on the LRS3 labeled data setting with the Conformer-BASE VSR model.

Effect of VSR Perceptual Loss. We experiment with four lip animation model variants: LAM-Baseline, LAM-LRS3-VSR-V, LAM-LRS3-VSR-L, LAM-LRS3-VSR-VL to generate 944 hours of synthetic data from Librispeech, respectively. The WER of the BASE model trained with only 438 hours of LRS3 data is 36.7% (see in Tab. 2), the results are shown in Tab. 3. Using 944 hours of synthetic data generated from the LAM-Baseline model, the WER improves to 32.2%. The individual visual VSR loss and linguistic VSR loss improve the WER to 31.7% and 31.4%, respectively. The linguistic VSR loss performs slightly better than the visual one. Further, the combination of visual and linguistic VSR loss further improves the WER to 30.8%, which indicates that the spatio-temporal visual and high-level semantic knowledge from a pre-trained VSR model could offer the lip animation model more accurate lip movements.

Lip animation model	WER (%)
LAM-Baseline	32.2
LAM-LRS3-VSR-V	31.7
LAM-LRS3-VSR-L	31.4
LAM-LRS3-VSR-VL	**30.8**

Table 3. Ablation studies on the proposed VSR perceptual loss.

Effect of Increasing Lip Sources Diversity. To analyze the impact of the diversity of lip sources we generate 944 hours of synthetic data generated from Librispeech using two target face sources: CelebA and LRS3 (pre-training). Images from LRS3 (pre-training) do not introduce any new lip images for VSR training. The results are reported in Tab. 4. Using the synthetic data with LRS3 lip images, the BASE model achieves the WER 32.1% which is 1.3% worse than that with CelebA lip images, which means the external lip images from CelebA offer better WER improvement.

Lip source	Unique images	WER (%)
LRS3-pretrain	5,090	32.1
CelebA	202,599	**30.8**

Table 4. Ablation studies on the impact of lip sources diversity. LRS3-pretrain indicates using the LRS3 (pre-training) images.

Effect of the Scale of Speech Corpus. Using synthetic data generated from TED-LIUM 3, Librispeech, and Common Voice with LAM-LRS3-VSR-VL model synthetic data, the LRS3 (test) WER improves from 36.7% to 30.9%, 30.8%, and 30.1%, respectively (see in Tab. 5). When combining all synthetic data together (3,652 hours) for training, the WER improves to 28.4%. We observe that richer speech corpora can offer larger performance improvements. Furthermore, we find that the improvement from scaling speech corpora is more pronounced than when adding VSR perceptual loss and lip sources diversity, suggesting that our proposed method benefits greatly from large speech datasets.

Training data	Hours	WER (%)
LRS3	438	36.7
LRS3 + TED-Synth	903	30.9
LRS3 + LBS-Synth	1,382	30.8
LRS3 + CV-Synth	2,681	30.1
LRS3 + [TED + LBS + CV]-Synth	4,090	**28.4**

Table 5. Ablation studies on different speech corpus. TED-Synth, LBS-Synth, CV-Synth indicates the synthetic data generated from TED-LIUM 3, Librispeech, and Common Voice, respectively.

Assessment of the Domain Mismatch Using VSR. We use 944 hours of Librispeech synthetic data generated from LAM-Baseline and LAM-LRS3-VSR-VL for VSR training. We conduct an assessment on the LRS3 real and synthetic test sets, the synthetic test set is obtained by passing the speech and the first frame in the real test set to the lip animation model. The results are illustrated in Fig. 3. By introducing the VSR perceptual loss, the WER on the synthetic test set is reduced from 106.8% to 49.7%, which indicates the VSR loss minimizes the domain gap between real and synthetic data. After training on synthetic data, the WER reduces for both real and synthetic data. Specifically, the WER is improved substantially from 106.8% to 36.0% and from 49.7% to 16.8% on synthetic test sets generated using LAM-Baseline and LAM-LRS3-VSR-VL, respectively. We observe that domain mismatch can be further reduced by combining real and synthetic data for VSR training.

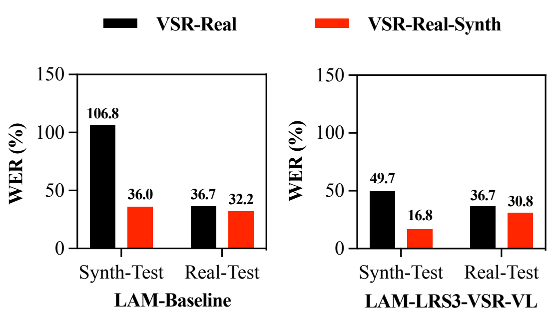

Figure 3. Assessment of the domain mismatch between real and synthetic video data using VSR models trained with only real data (VSR-Real, black) and real data together with synthetic data (VSR-Real-Synth, red), respectively. Synth-Test and Real-Test refer to the synthetic and real LRS3 (test) sets.

4.8. SynthVSR with TTS-generated Speech

In the real world, text corpora (e.g., Wikipedia) are more accessible than transcribed speech corpora. To explore the potential of using TTS-generated speech for SynthVSR, we use the off-the-shelf TTS model FastSpeech 2 [39] to generate 944 hours of synthetic speech from Librispeech transcriptions, which is further used to generate synthetic video clips using LAM-LRS3-VSR-VL model. Using these 944 hours of synthetic data to train the BASE model under the LRS3 labeled data setting, the WER is improved from 36.7% to 32.9%, as shown in Tab. 6. Although TTS-generated speech is not as natural as the original LibriSpeech data, the improvement of using synthetic data is still decent (3.8%), indicating the further potential of SynthVSR in real-world applications where labeled video data is sparse, such as the healthcare industry.

Training data	Hours	WER (%)
LRS3	438	36.7
LRS3 + LBS-Synth	1,382	30.8
LRS3 + TTS-LBS-Synth	1,382	32.9

Table 6. Experiments results of using the synthetic speech generated synthetic video (TTS-LBS-Synth) for VSR training.

5. Limitations & Societal Impact

The excellent performance improvement of SynthVSR comes at the cost of high computational demands during training, as we use additional large-scale synthetic video data. Also, although SynthVSR has achieved state-of-the-art VSR performance, the LRS3 test set is relatively limited (0.9 hours), which is from TED talks, while real-world videos are more challenging (e.g., egocentric videos). SynthVSR has many positive real-world applications, such as helping the hearing-impaired or people with aphonia with everyday communication.

6. Conclusion

We have presented a semi-supervised method for VSR enhanced with synthetic lip movements. The speech-driven lip animation model is proposed to generate synthetic video data from labeled speech datasets and face images for scaling up VSR. Our method achieves state-of-the-art results on LRS3, outperforming prior work trained on more labeled or unlabeled real video data. Our work fosters future research on generating and exploiting synthetic visual data for VSR.

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2, 6

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: A large-scale dataset for visual speech recognition. *arXiv:1809.00496*, 2018. 2, 3, 5

[3] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. ASR is all you need: Cross-modal distillation for lip reading. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2143–2147, 2020. 2, 3, 6

[4] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, 2020. 2, 5

[5] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. LipNet: Sentence-level lipreading. *arXiv:1611.01599*, 2016. 2, 3

[6] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460, 2020. 5

[7] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *European Conference on Computer Vision*, pages 35–51. Springer, 2020. 3

[8] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision*, pages 520–535, 2018. 3

[9] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? In *British Machine Vision Conference*, 2017. 3

[10] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 3, 5, 7

[11] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103. Springer, 2016. 2, 6

[12] Soo-Whan Chung, Hong Goo Kang, and Joon Son Chung. Seeing voices and hearing voices: Learning discriminative embeddings using cross-modal self-supervision. In *INTERSPEECH*, 2020. 3

[13] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3D human pose estimation: Motion to the rescue. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 3

[14] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv:1804.03619*, 2018. 3, 5, 6

[15] Amin Fazel, Wei Yang, Yulan Liu, Roberto Barra-Chicote, Yixiong Meng, Roland Maas, and Jasha Droppo. SynthASR: Unlocking synthetic data for speech recognition. *arXiv:2106.07803*, 2021. 3

[16] Chuang Gan, Jeremy Schwartz, Seth Alter, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, et al. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv:2007.04954*, 2020. 3

[17] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen 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. 2

[18] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. In *INTERSPEECH*, 2020. 2, 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[20] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *International Conference on Speech and Computer*, pages 198–208, 2018. 2, 5

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 5

[22] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018. 5

[23] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision*, December 2015. 2, 5

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv:1711.05101*, 2017. 5

[25] Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. Auto-AVSR: Audio-visual speech recognition with automatic labels. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023. 2, 5, 7

[26] Pingchuan Ma, Rodrigo Mira, Stavros Petridis, Björn W Schuller, and Maja Pantic. LiRA: Learning visual speech representations from audio through self-supervision. In *INTERSPEECH*, 2021. 2, 3, 6

[27] Pingchuan Ma, Stavros Petridis, and Maja Pantic. End-to-end audio-visual speech recognition with Conformers. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7613–7617, 2021. 2, 3, 6

[28] Pingchuan Ma, Stavros Petridis, and Maja Pantic. Visual speech recognition for multiple languages in the wild. *Nature Machine Intelligence*, 2022. 2, 3, 5, 6, 7

[29] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Contrastive learning of global and local audio-visual representations. *arXiv:2104.05418*, 2021. 3

[30] Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan. Recurrent neural network transducer for audio-visual

speech recognition. In *IEEE Automatic Speech Recognition and Understanding Workshop*, pages 905–912, 2019. 7

[31] Sepand Mavandadi, Tara N Sainath, Kevin Hu, and Zelin Wu. A deliberation-based joint acoustic and text decoder. In *INTERSPEECH*, 2021. 3

[32] Xichen Pan, Peiyu Chen, Yichen Gong, Helong Zhou, Xinbing Wang, and Zhouhan Lin. Leveraging uni-modal self-supervised learning for multimodal audio-visual speech recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022. 3

[33] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210, 2015. 2, 5

[34] Xingchao Peng, Baochen Sun, Karim Ali, and Kate Saenko. Learning deep object detectors from 3D models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1278–1286, 2015. 3

[35] KR Prajwal, Triantafyllos Afouras, and Andrew 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. 3, 7

[36] Aayush Prakash, Shaad Boochoon, Mark Brophy, David Acuna, Eric Cameracci, Gavriel State, Omer Shapira, and Stan Birchfield. Structured domain randomization: Bridging the reality gap by context-aware synthetic data. In *2019 International Conference on Robotics and Automation*, pages 7249–7255, 2019. 3

[37] Sucheng Ren, Yong Du, Jianming Lv, Guoqiang Han, and Shengfeng He. Learning from the master: Distilling cross-modal advanced knowledge for lip reading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13325–13333, 2021. 2, 6

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015. 5

[39] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv:2006.04558*, 2020. 8

[40] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3234–3243, 2016. 3

[41] Dmitriy Serdyuk, Otavio Braga, and Olivier Siohan. Transformer-based video front-ends for audio-visual speech recognition. In *INTERSPEECH*, 2022. 2, 3, 7

[42] Dmitriy Serdyuk, Olivier Siohan, and Otavio de Pinho Forin Braga. Audio-visual speech recognition is worth 32x32x8 voxels. In *IEEE Automatic Speech Recognition and Understanding Workshop*, 2021. 2, 3, 7

[43] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations*, 2022. 2, 3, 6, 7

[44] Brendan Shillingford, Yannis Assael, Matthew W Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, et al. Large-scale visual speech recognition. In *INTERSPEECH*, 2019. 2, 3, 6

[45] Themos Stafylakis and Georgios Tzimiropoulos. Combining residual networks with LSTMs for lipreading. *arXiv:1703.04105*, 2017. 3

[46] Jörgen Valk and Tanel Alumäe. Voxlingua107: A dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop*, pages 652–658, 2021. 5

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 3

[48] Yuri Viazovetskyi, Vladimir Ivashkin, and Evgeny Kashin. StyleGAN2 distillation for feed-forward image manipulation. In *European Conference on Computer Vision*, pages 170–186. Springer, 2020. 4

[49] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven facial animation with temporal GANs. In *British Machine Vision Conference*, 2018. 2, 3

[50] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with GANs. *International Journal of Computer Vision*, 128(5):1398–1413, 2020. 2, 3

[51] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12635–12644, 2020. 3

[52] Bo Xu, Cheng Lu, Yandong Guo, and Jacob Wang. Discriminative multi-modality speech recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14433–14442, 2020. 2, 6

[53] Xingxuan Zhang, Feng Cheng, and Shilin Wang. Spatio-temporal fusion based convolutional sequence learning for lip reading. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 713–722, 2019. 2

[54] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9299–9306, 2019. 3

[55] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4176–4186, 2021. 3