A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild

by

Prajwal K R, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, C V Jawahar

in

ACMM-2020 International Conference.

Report No: IIIT/TR/2020/-1

Centre for Visual Information Technology
International Institute of Information Technology
Hyderabad - 500 032, INDIA
August 2020

A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild

K R Prajwal*
prajwal.k@research.iiit.ac.in
IIIT, Hyderabad, India

Rudrabha Mukhopadhyay*
radrabha.m@research.iiit.ac.in
IIIT, Hyderabad, India

Vinay P. Namboodiri
vpn22@bath.ac.uk
University of Bath, England

C V Jawahar
jawahar@iiit.ac.in
IIIT, Hyderabad, India

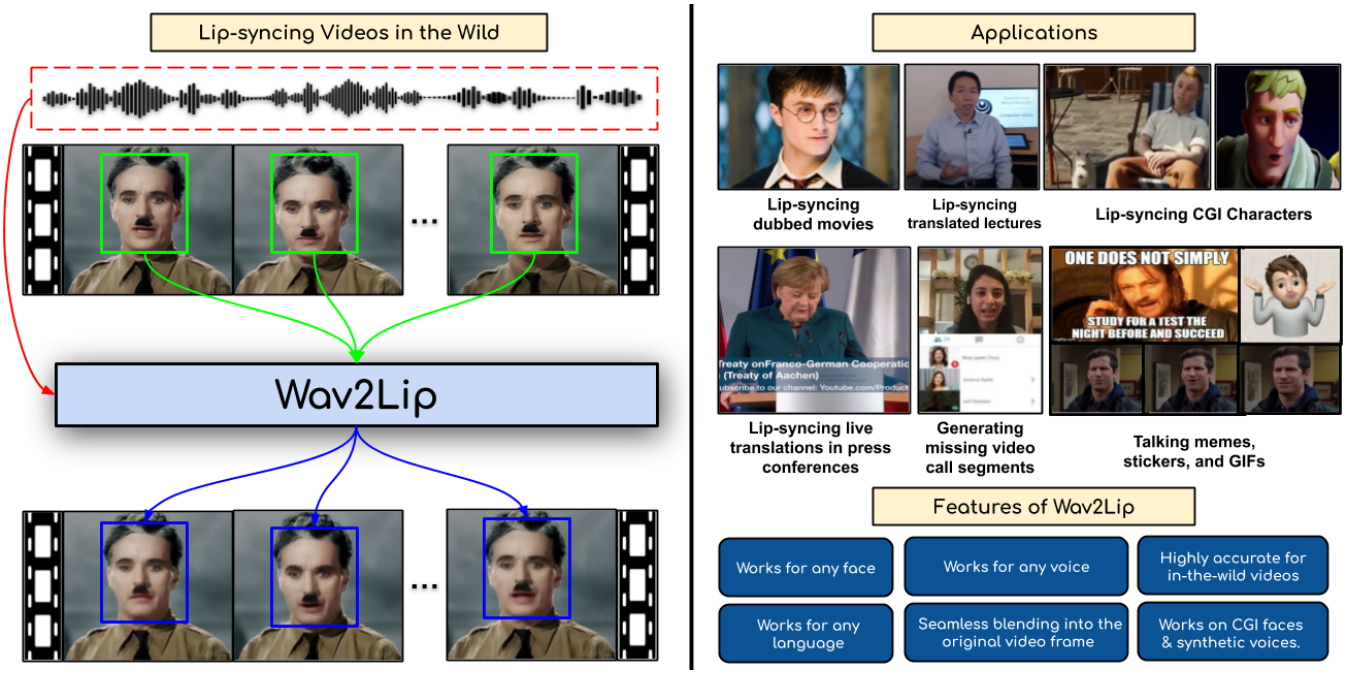

Figure 1: Our novel Wav2Lip model produces significantly more accurate lip-synchronization in dynamic, unconstrained talking face videos. Quantitative metrics indicate that the lip-sync in our generated videos are almost as good as real-synced videos. Thus, we believe that our model can enable a wide range of real-world applications where previous speaker-independent lip-syncing approaches [17, 18] struggle to produce satisfactory results.

ABSTRACT

In this work, we investigate the problem of lip-syncing a talking face video of an arbitrary identity to match a target speech segment. Current works excel at producing accurate lip movements on a static image or videos of specific people seen during the training phase. However, they fail to accurately morph the lip movements of arbitrary identities in dynamic, unconstrained talking face videos, resulting in significant parts of the video being out-of-sync with the new audio. We identify key reasons pertaining to this and hence resolve them by learning from a powerful lip-sync discriminator. Next, we propose new, rigorous evaluation benchmarks and metrics to accurately measure lip synchronization in unconstrained videos. Extensive quantitative evaluations on our challenging benchmarks show that the lip-sync accuracy of the videos generated by our Wav2Lip model is almost as good as real synced videos. We provide a demo video clearly showing the substantial impact of our Wav2Lip model and evaluation benchmarks on our website: cvit.iiit.ac.in/research/projects/cvit-projects/a-lip-sync-expert-is-all-you-need-for-speech-to-lip-generation-in-the-wild. The code and models are released here: github.com/Rudrabha/Wav2Lip. You

*Both of the authors have contributed equally to this research.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413532

can also try out the interactive demo at this link: bhaasha.iiit.ac.in/lipsync.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Learning from critiques**; *Phonology / morphology*.

KEYWORDS

lip sync;video generation;talking face generation

ACM Reference Format:

K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C V Jawahar. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3413532

1 INTRODUCTION

With the exponential rise in the consumption of audio-visual content [21], rapid video content creation has become a quintessential need. At the same time, making these videos accessible in different languages is also a key challenge. For instance, a deep learning lecture series, a famous movie, or a public address to the nation, if translated to desired target languages, can become accessible to millions of new viewers. A crucial aspect of translating such talking face videos or creating new ones is correcting the lip sync to match the desired target speech. Consequently, lip-syncing talking face videos to match a given input audio stream has received considerable attention [6, 13, 17, 18, 23] in the research community.

Initial works [19, 22] using deep learning in this space learned a mapping from speech representations to lip landmarks using several hours of a single speaker. More recent works [13, 23] in this line directly generate images from speech representations and show exceptional generation quality for specific speakers which they have been trained upon. Numerous practical applications, however, require models that can readily work for generic identities and speech inputs. This has led to the creation of speaker-independent speech to lip generation models [17, 18] that are trained on thousands of identities and voices. They can generate accurate lip motion on a single, static image of any identity in any voice, including that of a synthetic speech generated by a text-to-speech system [18]. However, to be used for applications like translating a lecture/TV series, for example, these models need to be able to morph the broad diversity of lip shapes present in these dynamic, unconstrained videos as well, and not just on static images.

Our work builds upon this latter class of speaker-independent works that aspire to lip-sync talking face videos of any identity and voice. We find that these models that work well for static images are unable to accurately morph the large variety of lip shapes in unconstrained video content, leading to significant portions of the generated video being out-of-sync with the new target audio. A viewer can recognize an out-of-sync video segment as small as just $\approx 0.05 - 0.1$ seconds [9] in duration. Thus, convincingly lip-syncing a real-world video to an entirely new speech is quite challenging, given the tiny degree of allowed error. Further, the fact that we are aiming for a speaker-independent approach without any additional speaker-specific data overhead makes our task even more difficult. Real-world videos contain rapid pose, scale, and illumination changes and the generated face result must also seamlessly blend into the original target video.

We start by inspecting the existing speaker-independent approaches for speech to lip generation. We find that these models do not adequately penalize wrong lip shapes, either as a result of using only reconstruction losses or weak lip-sync discriminators. We adapt a powerful lip-sync discriminator that can enforce the generator to consistently produce accurate, realistic lip motion. Next, we re-examine the current evaluation protocols and devise new, rigorous evaluation benchmarks derived from three standard test sets. We also propose reliable evaluation metrics using SyncNet [9] to precisely evaluate lip sync in unconstrained videos. We also collect and release ReSyncED, a challenging set of real-world videos that can benchmark how the models will perform in practice. We conduct extensive quantitative and subjective human evaluations and outperform previous methods by a large margin across all benchmarks. Our key contributions/claims are as follows:

- We propose a novel lip-synchronization network, Wav2Lip, that is significantly more accurate than previous works for lip-syncing arbitrary talking face videos in the wild with arbitrary speech.
- We propose a new evaluation framework, consisting of new benchmarks and metrics, to enable a fair judgment of lip synchronization in unconstrained videos.
- We collect and release ReSyncED, a Real-world lip-Sync Evaluation Dataset to benchmark the performance of the lip-sync models on completely unseen videos in the wild.
- Wav2Lip is the first speaker-independent model to generate videos with lip-sync accuracy that matches the real synced videos. Human evaluations indicate that the generated videos of Wav2Lip are preferred over existing methods and unsynced versions more than 90% of the time.

A demo video can be found on our website[1] with several qualitative examples that clearly illustrate the impact of our model. We will also release an interactive demo on the website allowing users to try out the model using audio and video samples of their choice. The rest of the paper is organized as follows: Section 2 surveys the recent developments in the area of speech to lip generation, Section 3 discusses the issues with the existing works and describes our proposed approach to mitigate them, Section 4 proposes a new, reliable evaluation framework. We describe the various potential applications and address some of the ethical concerns in Section 5 and conclude in Section 6.

2 RELATED WORK

2.1 Constrained Talking Face Generation from Speech

We first review works on talking face generation that are either constrained by the range of identities they can generate or the range of vocabulary they are limited to. Realistic generation of talking face videos was achieved by a few recent works [19, 22] on videos of Barack Obama. They learn a mapping between the input audio

[1] cvit.iiit.ac.in/research/projects/cvit-projects/a-lip-sync-expert-is-all-you-need-for-speech-to-lip-generation-in-the-wild

and the corresponding lip landmarks. As they are trained on only a specific speaker, they cannot synthesize for new identities or voices. They also require a large amount of data of a particular speaker, typically a few hours. A recent work along this line [13] proposes to seamlessly edit videos of individual speakers by adding or removing phrases from the speech. They still require an hour of data per speaker to achieve this task. Very recently, another work [23] tries to minimize this data overhead by using a two-stage approach, where they first learn speaker-independent features and then learn a rendering mapping with $\approx$ 5 minutes of data of the desired speaker. However, they train their speaker-independent network on a significantly smaller corpus and also have an additional overhead of requiring clean training data of each target speaker to generate for that speaker. Another limitation of existing works is in terms of the vocabulary. Several works [5, 26, 28] train on datasets with a limited set of words such as GRID [10] (56 words), TIMIT [14] and LRW [8] (1000 words) which significantly hampers a model from learning the vast diversity of phoneme-viseme mappings in real videos [18]. Our work focuses on lip-syncing unconstrained talking face videos to match any target speech, not limited by identities, voices, or vocabulary.

2.2 Unconstrained Talking Face Generation from Speech

Despite the rise in the number of works on speech-driven face generation, surprisingly, very few works have been designed to lip-sync videos of arbitrary identities, voices, and languages. They are not trained on a small set of identities or a small vocabulary. This allows them to, at test time, lip-sync random identities for any speech. To the best of our knowledge, only two such prominent works [17, 18] exist in the current literature. Note that [17] is an extended version of [7]. Both these works [17, 18] formulate the task of learning to lip-sync in the wild as follows: *Given a short speech segment S and a random reference face image R, the task of the network is to generate a lip-synced version L_g of the input face that matches the audio.* Additionally, the LipGAN model also inputs the target face with bottom-half masked to act as a pose prior. This was crucial as it allowed the generated face crops to be seamlessly pasted back into the original video without further post-processing. It also trains a discriminator in conjunction with the generator to discriminate in-sync or out-of-sync audio-video pairs. Both these works, however, suffer from a significant limitation: they work very well on static images of arbitrary identities but produce inaccurate lip generation when trying to lip-sync unconstrained videos in the wild. In contrast to the GAN setup used in LipGAN [18], we use a pre-trained, accurate lip-sync discriminator that is not trained further with the generator. We observe that this is an important design choice to achieve much better lip-sync results.

3 ACCURATE SPEECH-DRIVEN LIP-SYNCING FOR VIDEOS IN THE WILD

Our core architecture can be summed up as *"Generating accurate lip-sync by learning from a well-trained lip-sync expert"*. To understand this design choice, we first identify two key reasons why existing architectures (section 2.2) produce inaccurate lip-sync for videos in the wild. We argue that the loss functions, namely the L1 reconstruction loss used in both the existing works [17, 18] and the discriminator loss in LipGAN [18] are inadequate to penalize inaccurate lip-sync generation.

3.1 Pixel-level Reconstruction loss is a Weak Judge of Lip-sync

The face reconstruction loss is computed for the whole image, to ensure correct pose generation, preservation of identity, and even background around the face. The lip region corresponds to less than 4% of the total reconstruction loss (based on the spatial extent), so a lot of surrounding image reconstruction is first optimized before the network starts to perform fine-grained lip shape correction. This is further supported by the fact that the network begins morphing lips only at around half-way ($\approx 11^{th}$ epoch) through its training process ($\approx$ 20 epochs [18]). Thus, it is crucial to have an additional discriminator to judge lip-sync, as also done in LipGAN [18]. But, how powerful is the discriminator employed in LipGAN?

3.2 A Weak Lip-sync Discriminator

We find that the LipGAN's lip-sync discriminator is only about 56% accurate while detecting off-sync audio-lip pairs on the LRS2 test set. For comparison, the expert discriminator that we will use in this work is 91% accurate on the same test set. We hypothesize two major reasons for this difference. Firstly, LipGAN's discriminator uses a single frame to check for lip-sync. In Table 3, we show that a small temporal context is very helpful while detecting lip-sync. Secondly, the generated images during training contain a lot of artifacts due to the large scale and pose variations. We argue that training the discriminator in a GAN setup on these noisy generated images, as done in LipGAN, results in the discriminator focusing on the visual artifacts instead of the audio-lip correspondence. This leads to a large drop in off-sync detection accuracy (Table 3). We argue and show that the "real", accurate concept of lip-sync captured from the actual video frames can be used to accurately discriminate and enforce lip-sync in the generated images.

3.3 A Lip-sync Expert Is All You Need

Based on the above two findings, we propose to use a pre-trained expert lip-sync discriminator that is accurate in detecting sync in real videos. Also, it should not be fine-tuned further on the generated frames like it is done in LipGAN. One such network that has been used to correct lip-sync errors for creating large lip-sync datasets [1, 3] is the SyncNet [9] model. We propose to adapt and train a modified version of SyncNet [9] for our task.

3.3.1 Overview of SyncNet. SyncNet [9] inputs a window V of T_v consecutive face frames (lower half only) and a speech segment S of size $T_a \times D$, where T_v and T_a are the video and audio time-steps respectively. It is trained to discriminate sync between audio and video by randomly sampling an audio window $T_a \times D$ that is either aligned with the video (in-sync) or from a different time-step (out-of-sync). It contains a face encoder and an audio encoder, both comprising of a stack of 2D-convolutions. L2 distance is computed between the embeddings generated from these encoders, and the model is trained with a max-margin loss to minimize (or maximize) the distance between synced (or unsynced) pairs.

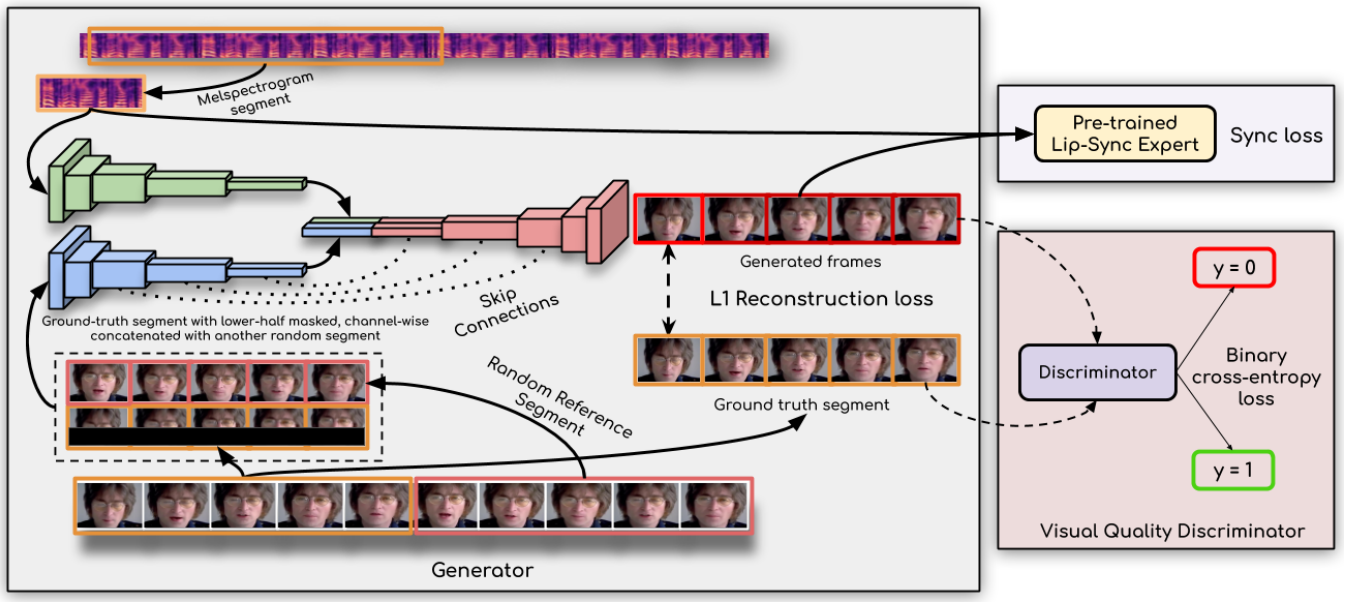

Figure 2: Our approach generates accurate lip-sync by learning from an "already well-trained lip-sync expert". Unlike previous works that employ only a reconstruction loss [17] or train a discriminator in a GAN setup [18], we use a pre-trained discriminator that is already quite accurate at detecting lip-sync errors. We show that fine-tuning it further on the noisy generated faces hampers the discriminator's ability to measure lip-sync, thus also affecting the generated lip shapes. Additionally, we also employ a visual quality discriminator to improve the visual quality along with the sync accuracy.

3.3.2 Our expert lip-sync discriminator. We make the following changes to SyncNet [9] to train an expert lip-sync discriminator that suits our lip generation task. Firstly, instead of feeding gray-scale images concatenated channel-wise as in the original model, we feed color images. Secondly, our model is significantly deeper, with residual skip connections [15]. Thirdly, inspired by this public implementation[2], we use a different loss function: cosine-similarity with binary cross-entropy loss. That is, we compute a dot product between the ReLU-activated video and speech embeddings v, s to yield a single value between $[0, 1]$ for each sample that indicates the probability that the input audio-video pair is in sync:

$$P_{\text{sync}} = \frac{v \cdot s}{max(\|v\|_2 \cdot \|s\|_2, \epsilon)} \tag{1}$$

We train our expert lip-sync discriminator on the LRS2 train split ($\approx$ 29 hours) with a batch size of 64, with $T_v = 5$ frames using the Adam optimizer [12] with an initial learning rate of $1e^{-3}$. Our expert lip-sync discriminator is about 91% accurate on the LRS2 test set, while the discriminator used in LipGAN is only 56% accurate on the same test set.

3.4 Generating Accurate Lip-sync by learning from a Lip-sync Expert

Now that we have an accurate lip-sync discriminator, we can now use it to penalize the generator (Figure 2) for inaccurate generation during the training time. We start by describing the generator architecture.

[2]github.com/joonson/syncnet_trainer

3.4.1 Generator Architecture Details. We use a similar generator architecture as used by LipGAN [18]. Our key contribution lies in training this with the expert discriminator. The generator G contains three blocks: (i) Identity Encoder, (ii) Speech Encoder, and a (iii) Face Decoder. The Identity Encoder is a stack of residual convolutional layers that encode a random reference frame R, concatenated with a pose-prior P (target-face with lower-half masked) along the channel axis. The Speech Encoder is also a stack of 2D-convolutions to encode the input speech segment S which is then concatenated with the face representation. The decoder is also a stack of convolutional layers, along with transpose convolutions for upsampling. The generator is trained to minimize L1 reconstruction loss between the generated frames L_g and ground-truth frames L_G:

$$L_{\text{recon}} = \frac{1}{N} \sum_{i=1}^{N} ||L_g - L_G||_1 \tag{2}$$

Thus, the generator is similar to the previous works, a 2D-CNN encoder-decoder network that generates each frame independently. How do we then employ our pre-trained expert lip-sync discriminator that needs a temporal window of $T_v = 5$ frames as input?

3.4.2 Penalizing Inaccurate Lip Generation. During training, as the expert discriminator trained in section 3.3 processes $T_v = 5$ contiguous frames at a time, we would also need the generator G to generate all the $T_v = 5$ frames. We sample a random contiguous window for the reference frames, to ensure as much temporal consistency of pose, etc. across the T_v window. As our generator processes each frame independently, we stack the time-steps along

the batch dimension while feeding the reference frames to get an input shape of $(N \cdot T_v, H, W, 3)$, where N, H, W are the batch-size, height, and width respectively. While feeding the generated frames to the expert discriminator, the time-steps are concatenated along the channel-dimension as was also done during the training of the discriminator. The resulting input shape to the expert discriminator is $(N, H/2, W, 3 \cdot T_v)$, where only the lower half of the generated face is used for discrimination. The generator is also trained to minimize the "expert sync-loss" E_{sync} from the expert discriminator:

$$E_{\text{sync}} = \frac{1}{N}\sum_{i=1}^{N} -\log(P^i_{\text{sync}}) \tag{3}$$

where P^i_{sync} is calculated according to Equation 1. Note that the expert discriminator's weights remain frozen during the training of the generator. This strong discrimination based purely on the lip-sync concept learned from real videos forces the generator to also achieve realistic lip-sync to minimize the lip-sync loss E_{sync}.

3.5 Generating Photo-realistic Faces

In our experiments, we observed that using a strong lip-sync discriminator forces the generator to produce accurate lip shapes. However, it sometimes results in the morphed regions to be slightly blurry or contain slight artifacts. To mitigate this minor loss in quality, we train a simple visual quality discriminator in a GAN setup along with the generator. Thus, we have two discriminators, one for sync accuracy and another for better visual quality. The lip-sync discriminator is *not* trained in a GAN setup for reasons explained in 3.2. On the other hand, since the visual quality discriminator does not perform any checks on lip-sync and only penalizes unrealistic face generations, it is trained on the generated faces.

The discriminator D consists of a stack of convolutional blocks. Each block consists of a convolutional layer followed by a Leaky ReLU activation [20]. The discriminator is trained to maximize the objective function L_{disc} (Equation 5):

$$L_{gen} = \mathbb{E}_{x \sim L_g}[log(1 - D(x)] \tag{4}$$

$$L_{disc} = \mathbb{E}_{x \sim L_G}[log(D(x))] + L_{gen} \tag{5}$$

where L_g corresponds to the images from the generator G, and L_G corresponds to the real images.

The generator minimizes Equation 6, which is the weighted sum of the reconstruction loss (Equation 2), the synchronization loss (Equation 3) and the adversarial loss L_{gen} (Equation 4):

$$L_{\text{total}} = (1 - s_w - s_g) \cdot L_{\text{recon}} + s_w \cdot E_{\text{sync}} + s_g \cdot L_{gen} \tag{6}$$

where s_w is the synchronization penalty weight, s_g is the adversarial loss which are empirically set to 0.03 and 0.07 in all our experiments. Thus, our complete network is optimized for both superior sync-accuracy and quality using two disjoint discriminators.

We train our model only on the LRS2 train set [1], with a batch size of 80. We use the Adam optimizer [12] with an initial learning rate of $1e^{-4}$ and betas $\beta_1 = 0.5, \beta_2 = 0.999$ for both the generator and visual quality discriminator D. Note that the lip-sync discriminator is not fine-tuned further, so its weights are frozen. We conclude the description of our proposed architecture by explaining how it works during the inference on real videos. Similar to LipGAN [18], the model generates a talking face video frame-by-frame. The visual input at each time-step is the current face crop (from the source frame), concatenated with the same current face crop with lower-half masked to be used as a pose prior. Thus, during inference, the model does not need to change the pose, significantly reducing artifacts. The corresponding audio segment is also given as input to the speech sub-network, and the network generates the input face crop, but with the mouth region morphed.

All our code and models will be released publicly. We will now quantitatively evaluate our novel approach against previous models.

4 QUANTITATIVE EVALUATION

Despite training only on the LRS2 train set, we evaluate our model across 3 different datasets. But before doing so, we re-investigate the current evaluation framework followed in prior works and why it is far from being an ideal way to evaluate works in this space.

4.1 Re-thinking the Evaluation Framework for Speech-driven Lip-Syncing in the Wild

The current evaluation framework for speaker-independent lip-syncing judges the models differently from how it is used while lip-syncing a real video. Specifically, instead of feeding the current frame as a reference (as described in the previous section), a random frame in the video is chosen as the reference to not leak the correct lip information during evaluation. We strongly argue that the evaluation framework in the previous paragraph is not ideal for evaluating the lip-sync quality and accuracy. Upon a closer examination of the above-mentioned evaluation system, we observed a few key limitations, which we discuss below.

4.1.1 Does not reflect the real-world usage. As discussed before, during generation at test time, the model must not change the pose, as the generated face needs to be seamlessly pasted into the frame. However, the current evaluation framework feeds random reference frames in the input, thus demanding the network to change the pose. Thus, the above system does not evaluate how the model would be used in the real world.

4.1.2 Inconsistent evaluation. As the reference frames are chosen at random, this means the test data is not consistent across different works. This would lead to an unfair comparison and hinder the reproducibility of results.

4.1.3 Does not support checking for temporal consistency. As the reference frames are randomly chosen at each time-step, temporal consistency is already lost as the frames are generated at random poses and scales. The current framework cannot support a new metric or a future method that aims to study the temporal consistency aspect of this problem.

4.1.4 Current metrics are not specific to lip-sync. The existing metrics, such as SSIM [27] and PSNR, were developed to evaluate overall image quality and not fine-grained lip-sync errors. Although LMD [4] focuses on the lip region, we found that lip landmarks can be quite inaccurate on generated faces. Thus, there is a need for a metric that is designed specifically for measuring lip-sync errors.

	LRW [8]			LRS2 [1]			LRS3 [3]		
Method	**LSE-D ↓**	**LSE-C ↑**	**FID ↓**	**LSE-D ↓**	**LSE-C ↑**	**FID ↓**	**LSE-D ↓**	**LSE-C ↑**	**FID ↓**
Speech2Vid [17]	13.14	1.762	11.15	14.23	1.587	12.32	13.97	1.681	11.91
LipGAN [18]	10.05	3.350	2.833	10.33	3.199	4.861	10.65	3.193	4.732
Wav2Lip (ours)	**6.512**	**7.490**	3.189	**6.386**	**7.789**	4.887	**6.652**	**7.887**	4.844
Wav2Lip + GAN (ours)	6.774	7.263	**2.475**	6.469	7.781	**4.446**	6.986	7.574	**4.350**
Real Videos	7.012	6.931	—	6.736	7.838	—	6.956	7.592	—

Table 1: We propose two new metrics "Lip-Sync Error-Distance" (lower is better) and "Lip-Sync Error-Confidence" (higher is better), that can reliably measure the lip-sync accuracy in unconstrained videos. We see that the lip-sync accuracy of the videos generated using Wav2Lip is almost as good as real synced videos. Note that we only train on the train set on LRS2 [1], but we comfortably generalize across all datasets without any further fine-tuning. We also report the FID score (lower is better), which clearly shows that using a visual quality discriminator improves the quality by a significant margin.

4.2 A Novel Benchmark and Metric for Evaluating Lip-Sync in the Wild

The reason for sampling random frames for evaluation is because, the current frame is already in sync with the speech, leading to leakage of lip-shape in the input itself. And previous works have not tried sampling different speech segments instead of sampling a different frame, as the ground-truth lip shape for the sampled speech is unavailable.

4.2.1 A Metric to Measure the Lip-Sync Error. We propose to use the pre-trained SyncNet [9] available publicly[3] to measure the lip-sync error between the generated frames and the randomly chosen speech segment. The accuracy of SyncNet averaged over a video clip is over 99% [9]. Thus, we believe this can be a good automatic evaluation method that explicitly tests for accurate lip-sync in unconstrained videos in the wild. Note that this is not the expert lip-sync discriminator that we have trained above, but the one released by Chung and Zisserman [9], which was trained on a different, non-public dataset. Using a SyncNet resolves major issues of the existing evaluation framework. We no longer need to sample random, temporally incoherent frames and SyncNet also takes into account short-range temporal consistency while evaluating lip-sync. Thus, we propose two new metrics automatically determined using the SyncNet model. The first is the average error measure calculated in terms of the distance between the lip and audio representations, which we code-name as "LSE-D" ("Lip Sync Error - Distance"). A lower LSE-D denotes a higher audio-visual match, i.e., the speech and lip movements are in sync. The second metric is the average confidence score, which we code-name as "LSE-C" (Lip Sync Error - Confidence). Higher the confidence, the better the audio-video correlation. A lower confidence score denotes that there are several portions of the video with completely out-of-sync lip movements. Further details can be found in the SyncNet paper [9].

4.2.2 A Consistent Benchmark to Evaluate Lip-sync in the wild. Now that we have an automatic, reliable metric that can be computed for any video and audio pairs, we can sample random speech samples instead of a random frame at each time-step. Thus, we can create a list of pairs of video and a pseudo-randomly chosen audio as a consistent test set. We create three consistent benchmarks test sets, one each using the test set videos of LRS2 [1], LRW [8], and LRS3 [3] respectively. For each video V_s, we take the audio from another randomly-sampled video V_t with the condition that the length of the speech V_t be less than V_s. We create $14K$ audio-video pairs using LRS2. Using the LRW test set, we create $28K$ pairs, and this set measures the performance on frontal/near-frontal videos [2]. We also create $14K$ pairs using the LRS3 test set, which will be a benchmark for lip-syncing in profile views as well. The complete evaluation toolkit will be publicly released for consistent and reliable benchmarking of lip-syncing videos in the wild.

[3] github.com/joonson/syncnet_python

4.3 Comparing the Models on the New Benchmark

We compare the previous two approaches [17, 18] on our newly created test set using the LSE-D and LSE-C metrics. During inference, we now feed the same reference and pose-prior at each time-step, similar to how it has been described before in the architecture section. The mean LSE-D and LSE-C scores are shown in Table 1 for the audio-video pairs in all three test splits. Additionally, to measure the quality of the generated faces, we also report the FrÃĭchet Inception Distance (FID). Our method outperforms previous approaches by a large margin indicating the significant effect of strong lip-sync discrimination. We can also see the significant improvement in quality after using a visual quality discriminator along with a lip-sync expert discriminator. However, we observe a minor drop in sync accuracy after using the visual quality discriminator. Thus, we will release both of these models, as they have a slight trade-off between visual quality and sync accuracy.

4.4 Real-World Evaluation

Apart from evaluating on just the standard datasets, our new evaluation framework and metrics allow us to evaluate on real-world videos on which these models are most likely to be used. Further, given the sensitivity of humans to audio-lip synchronization [9], it is necessary to also evaluate our results with the help of human evaluators. Thus, contrary to the previous works on speaker-independent lip-syncing, we conduct both quantitative and human evaluation experiments on unconstrained real videos from the web for the first time. Thus, we collect and publicly release "ReSyncED" a "Real-world Evaluation Dataset" to subjectively and objectively benchmark the performance of lip-sync works.

4.4.1 Curating ReSyncED. All our videos are downloaded from YouTube. We specifically choose three types of video examples. The first type "Dubbed", contains videos where the audio is naturally out-of-sync, such as dubbed movie clips or public addresses that are live translated to a different language (so the addresser's lips

Method	Video Type	LSE-D ↓	LSE-C ↑	FID ↓	Sync Acc.	Visual Qual.	Overall Experience	Preference
Unsynced Orig. Videos	Dubbed	12.63	0.896	—	0.21	4.81	3.07	3.15%
Speech2Vid [17]		14.76	1.121	19.31	1.14	0.93	0.84	0.00%
LipGAN [18]		10.61	2.857	12.87	2.98	3.91	3.45	2.35%
Wav2Lip (ours)		**6.843**	**7.265**	15.65	**4.13**	3.87	4.04	34.3%
Wav2Lip + GAN (ours)		7.318	6.851	**11.84**	4.08	**4.12**	**4.13**	**60.2%**
Without Lip-syncing	Random	17.12	2.014	—	0.15	4.56	2.98	3.24%
Speech2Vid [17]		15.22	1.086	19.98	0.87	0.79	0.73	0.00%
LipGAN [18]		11.01	3.341	14.60	3.42	3.77	3.57	3.16%
Wav2Lip (ours)		**6.691**	**8.220**	14.47	**4.24**	3.68	4.01	29.1%
Wav2Lip + GAN (ours)		7.066	8.011	**13.12**	4.18	**4.05**	**4.15**	**64.5%**
Without Lip-syncing	TTS	16.89	2.557	—	0.11	4.67	3.32	8.32%
Speech2Vid [17]		14.39	1.471	17.96	0.76	0.71	0.69	0.00%
LipGAN [18]		10.90	3.279	11.91	2.87	3.69	3.14	1.64%
Wav2Lip (ours)		**6.659**	**8.126**	12.77	**3.98**	3.87	3.92	41.2%
Wav2Lip + GAN (ours)		7.225	7.651	**11.15**	3.85	**4.13**	**4.05**	**51.2%**
Untranslated Videos		7.767	7.047	—	4.83	4.91	—	—

Table 2: Real world evaluation using our newly collected ReSyncED benchmark. We evaluate using both quantitative metrics and human evaluation scores across three classes of real videos. We can see that in all cases, the Wav2Lip model produces high-quality, accurate lip-syncing videos. Specifically, the metrics indicate that our lip-synced videos are as good as the real synced videos. We also note that human evaluations indicate that there is a scope for improvement when trying to lip-sync TTS generated speech. Finally, it is worth noting that our lip-synced videos are preferred over existing methods or the actual unsynced videos over 90% of the time.

are out-of-sync with the translated speech). The second type is "Random", where we have a collection of videos and we create random audio-visual pairs similar to 4.2.2. The third and final type of videos, "TTS", has been specifically chosen for benchmarking the lip-syncing performance on synthetic speech obtained from a text-to-speech system. This is essential for future works that aspire to automatically translate videos (Face-to-Face Translation [18]) or rapidly create new video content. We manually transcribe the text, use Google Translate (about 5 languages totally) and publicly available text-to-speech models to generate synthetic translated speech for the videos in this category. The task is to correct lip movements in the original videos to match this synthetic speech.

4.4.2 Real-world Evaluation on ReSyncED. We first evaluate the generated real video results using our new automatic metrics, "LSE-D" and "LSE-C" obtained from SyncNet [9]. For the human evaluation, we ask 14 evaluators to judge the different synced versions of the videos based on the following parameters: (a) Sync Accuracy (b) Visual Quality (to evaluate the extent of visual artifacts), (c) Overall Experience (to evaluate the overall experience of the audio-visual content), and (d) Preference, where the viewer chooses the version of the video that is most appealing to watch. The first three parameters are scored between $1-5$, and (d) is a single-choice voting, and we report the percentage of votes obtained by a model. We evaluate each of the three classes of videos separately and report our results in Table 2. An outcome worth noting is that the previous works [17, 18] which produce several out-of-sync segments are less preferred over the unsynced version as the latter still preserves good Visual quality. Thus, ours is the first work that provides a significant improvement over unsynced talking face videos in-the-wild. We also show some qualitative comparisons in Figure 3 which contains a few generated samples from the ReSyncED test set.

4.5 Is our expert discriminator best among the alternatives?

Model	Fine-tuned?	Off-sync Acc.	LSE-D	LSE-C
$T_v = 1$ [18]	✓	55.6%	10.33	3.19
Ours, $T_v = 1$	×	79.3%	8.583	4.845
Ours, $T_v = 3$	✓	72.3%	10.14	3.214
Ours, $T_v = 3$	×	87.4%	7.230	6.533
Ours, $T_v = 5$	✓	73.6%	9.953	3.508
Ours, $T_v = 5$	×	**91.6%**	**6.386**	**7.789**

Table 3: A larger temporal window allows for better lip-sync discrimination. On the other hand, training the lip-sync discriminator on the generated faces deteriorates its ability to detect off-sync audio-lip pairs. Consequently, training a lip-sync generator using such a discriminator leads to poorly lip-synced videos.

Our expert discriminator uses $T_v = 5$ video frames to measure the lip-sync error. It is also not fine-tuned on the generated faces in a GAN setup. We justify these two design choices in this ablation study. We can test the discriminator's performance by randomly sampling in-sync and off-sync pairs from the LRS2 test set. We vary the size of $T_v = 1, 3, 5$ to understand its effect on detecting sync. We also fine-tune/freeze each of the three variants of T_v while training the Wav2Lip model. Thus, we get a total of 6 variations in Table 3 from which we can clearly make two observations. Increasing the temporal window size T_v consistently provides a better lip-sync discrimination performance. More importantly, we see that if we fine-tune the discriminator on the generated faces that contain artifacts, then the discriminator loses its ability to detect out-of-sync audio-visual pairs. We argue that this happens because the fine-tuned discriminator focuses on the visual artifacts in the generated

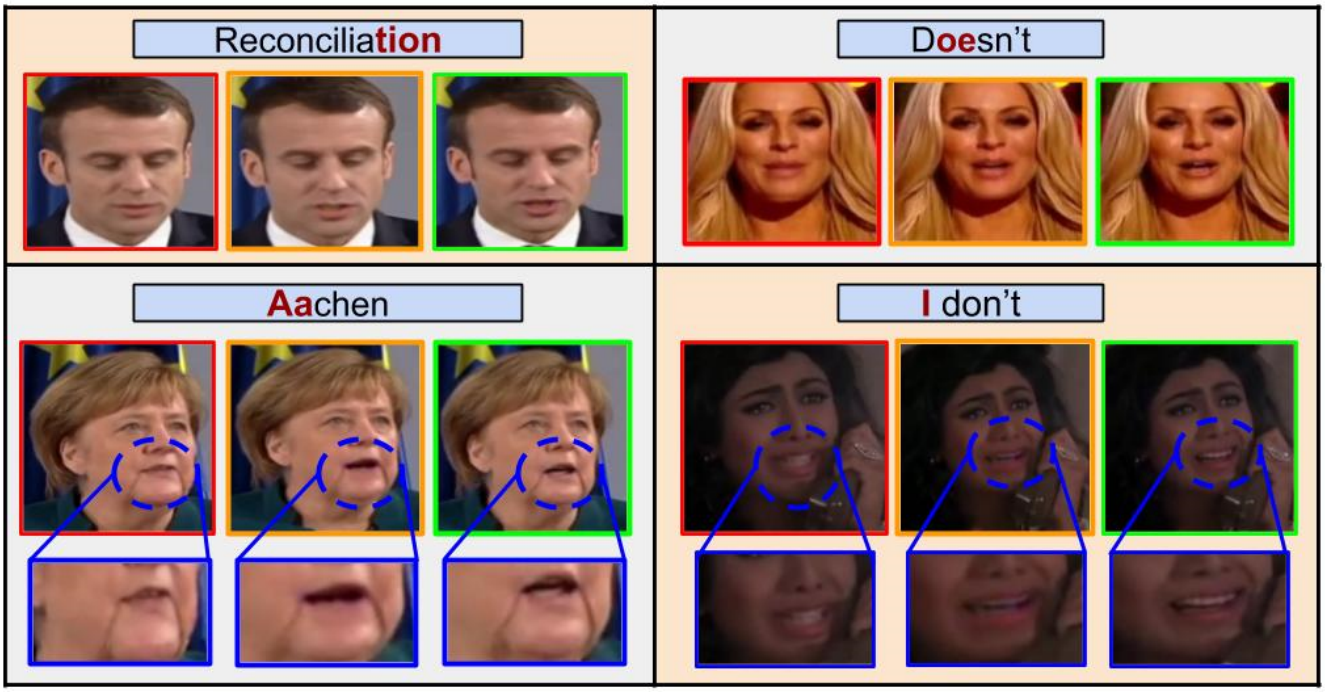

Figure 3: Examples of faces generated from our proposed models (green and yellow outlines). We compare with the current best approach [18] (red outline). The text is shown for illustration to denote the utterance being spoken in the frame shown. We can see that our model produces accurate, natural lip shapes. The addition of a visual quality discriminator also significantly improves the visual quality. We strongly encourage the reader to check out the demo video on our website.

faces for discrimination, rather than the fine-grained audio-lip correspondence. Thus, it classifies the *real* unsynced pairs as "in-sync", since these real face images do not contain any artifacts. Down the line, using such a weak discriminator leads to poor lip-sync penalization for our generator, resulting in poorly lip-synced talking face videos.

5 APPLICATIONS & FAIR USE

At a time when our content consumption and social communication is becoming increasingly audio-visual, there is a dire need for large-scale video translation and creation. Wav2Lip can play a vital role in fulfilling these needs, as it is accurate for videos in the wild. For instance, online lecture videos that are typically in English can now be lip-synced to (automatically) dubbed speech in other local languages (Table 2, last block). We can also lip-sync dubbed movies making them pleasant to watch (Table 2, first block). Every day throughout the globe, press conferences and public addresses are live translated but the addresser's lips are out of sync with the translated speech. Our model can seamlessly correct this. Automatically animating the lips of CGI characters to the voice actors' speech can save several hours of manual effort while creating animated movies and rich, conversational game content. We demonstrate our model on all these applications and more in the demo video on our website.

We believe that it is also essential to discuss and promote fair use of the increasingly capable lip-sync works. The vast applicability of our models with near-realistic lip-syncing capabilities for any identity and voice raises concerns about the potential for misuse. Thus, we strongly suggest that any result created using our code and models must unambiguously present itself as synthetic. In addition to the strong positive impact mentioned above, our intention to completely open-source our work is that it can simultaneously also encourage efforts [11, 16, 24, 25] in detecting manipulated video content and their misuse. We believe that Wav2Lip can enable several positive applications and also encourage productive discussions and research efforts regarding fair use of synthetic content.

6 CONCLUSION

In this work, we proposed a novel approach to generate accurate lip-synced videos in the wild. We have highlighted two major reasons why current approaches are inaccurate while lip-syncing unconstrained talking face videos. Based on this, we argued that a pre-trained, accurate lip-sync "expert" can enforce accurate, natural lip motion generation. Before evaluating our model, we re-examined the current quantitative evaluation framework and highlight several major issues. To resolve them, we proposed several new evaluation benchmarks and metrics, and also a real-world evaluation set. We believe future works can be reliably judged in this new framework. Our Wav2Lip model outperforms the current approaches by a large margin in both quantitative metrics and human evaluations. We also investigated the reasons behind our design choices in the discriminator in an ablation study. We encourage the readers to view the demo video on our website. We believe our efforts and ideas

in this problem can lead to new directions such as synthesizing expressions and head-poses along with the accurate lip movements.

REFERENCES

[1] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. 2018. Deep Audio-Visual Speech Recognition. In *arXiv:1809.02108*.

[2] T. Afouras, J. S. Chung, and A. Zisserman. 2018. The Conversation: Deep Audio-Visual Speech Enhancement. In *INTERSPEECH*.

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

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

[5] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7832–7841.

[6] Lele Chen, Haitian Zheng, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 2019. Sound to Visual: Hierarchical Cross-Modal Talking Face Video Generation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshops*.

[7] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. 2017. You said that? *arXiv preprint arXiv:1705.02966* (2017).

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

[9] Joon Son Chung and Andrew Zisserman. 2016. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*.

[10] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America* 120, 5 (2006), 2421–2424.

[11] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The DeepFake Detection Challenge Dataset. arXiv:2006.07397 [cs.CV]

[12] John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research* 12, 7 (2011).

[13] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. 2019. Text-based editing of talking-head video. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–14.

[14] Naomi Harte and Eoin Gillen. 2015. TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia* 17, 5 (2015), 603–615.

[15] 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.

[16] Chih-Chung Hsu, Yi-Xiu Zhuang, and Chia-Yen Lee. 2020. Deep Fake Image Detection based on Pairwise Learning. *Applied Sciences* 10 (2020), 370.

[17] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. 2019. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision* 127, 11-12 (2019), 1767–1779.

[18] Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. 2019. Towards Automatic Face-to-Face Translation. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 1428–1436.

[19] Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brébisson, and Yoshua Bengio. 2017. Obamanet: Photo-realistic lip-sync from text. *arXiv preprint arXiv:1801.01442* (2017).

[20] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, Vol. 30. 3.

[21] NPD. 2016. *52 Percent of Millennial Smartphone Owners Use their Device for Video Calling, According to The NPD Group*. https://www.npd.com/wps/portal/npd/us/news/press-releases/2016/52-percent-of-millennial-smartphone-owners-use-their-device-for-video-calling-according-to-the-npd-group/

[22] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 95.

[23] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. 2019. Neural Voice Puppetry: Audio-driven Facial Reenactment. *arXiv preprint arXiv:1912.05566* (2019).

[24] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. 2020. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. arXiv:2001.00179 [cs.CV]

[25] Eleanor Tursman, Marilyn George, Seny Kamara, and James Tompkin. 2020. Towards Untrusted Social Video Verification to Combat Deepfakes via Face Geometry Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[26] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2019. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision* (2019), 1–16.

[27] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[28] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2018. Talking Face Generation by Adversarially Disentangled Audio-Visual Representation. *arXiv preprint arXiv:1807.07860* (2018).