Towards Generating Ultra-High Resolution Talking-Face Videos with Lip synchronization

Anchit Gupta, IIIT-Hyderabad
Rudrabha Mukhopadhyay, IIIT-Hyderabad
Sindhu Balachandra, IIIT-Hyderabad
Faizan Farooq Khan, IIIT-Hyderabad
Vinay P. Namboodiri, University of Bath
C. V. Jawahar, IIIT-Hyderabad

{anchit.gupta,radrabha.m,sindhu.hegde}@research.iiit.ac.in, faizan.farooq@students.iiit.ac.in,
vpn22@bath.ac.uk, jawahar@iiit.ac.in

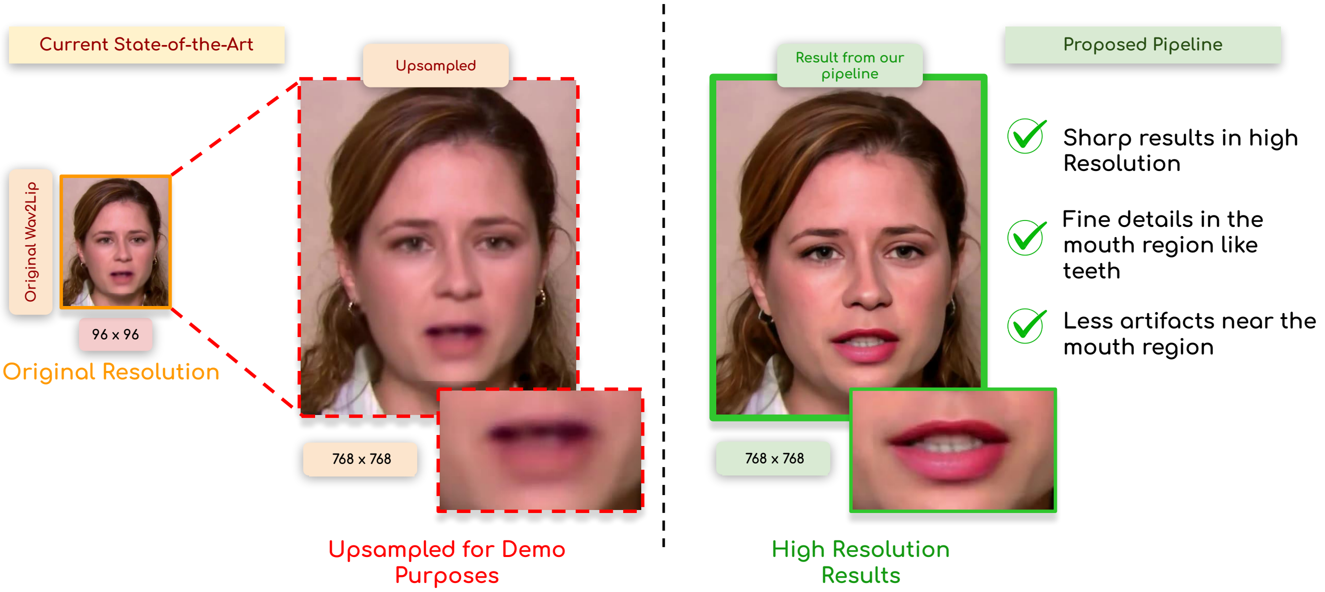

Figure 1: We propose the first talking-face generation network, which can lip-sync any identity at ultra-high resolutions like 4K. Our model captures fine-grained details of the lip region, including color, texture, and essential features like teeth. While the current state-of-the-art model Wav2Lip [16] generates faces at 96 × 96 pixels (left part), our proposed method synthesizes 64 times more pixels, rendering realistic, high-quality results at 768 × 768 pixels.

Abstract

Talking-face video generation works have achieved state-of-the-art results in synthesizing videos with lip synchronization. However, most of the previous works deal with low-resolution talking-face videos (up to 256×256 pixels), thus, generating extremely high-resolution videos still remains a challenge. We take a giant leap in this work and propose a novel method to synthesize talking-face videos at resolutions as high as 4K! Our task presents several key challenges: (i) Scaling the existing methods to such high resolutions is resource-constrained, both in terms of compute and the availability of very high-resolution datasets, (ii) The synthesized videos need to be spatially and temporally coherent. The sheer number of pixels that the model needs to generate while maintaining the temporal consistency at the video level makes this task non-trivial and has never been attempted before in literature. To address these issues, we propose to train the lip-sync generator in a compact Vector Quantized (VQ) space for the first time. Our *core idea to encode the faces in a compact 16×16 representation allows us to model high-resolution videos. In our framework, we learn the lip movements in the quantized space on the newly collected 4K Talking Faces (4KTF) dataset. Our approach is speaker agnostic and can handle various languages and voices. We benchmark our technique against several competitive works and show that we can achieve a remarkable 64-times more pixels than the current state-of-the-art! Our supplementary demo video depicts additional qualitative results, comparisons, and several real-world applications, like professional movie editing enabled by our model.*

1. Introduction

When was the last time we watched a video? For many of us, it will be well within 24 hours! In fact, for the majority of people, videos are the most common form of entertainment [1]. The rise of streaming platforms like YouTube and Over-The-Top (OTT) media platforms like Netflix has made video production more accessible to the masses. Such is the impact that over 200 thousand minutes of videos are streamed solely on Netflix every day! Video conferencing is yet another area that has seen a massive influx of users. According to a recent report [2], a video conferencing platform like Zoom enables over 300 million daily meetings amounting to 3.3 trillion minutes per year! Recently, due to the COVID-19 pandemic, the need for online lectures is gaining tremendous user attention. News reading, video calling, vlogs, marketing videos, and often a large part of movie scenes contain videos of the speaker. These videos are termed "talking-face videos". As the overall video content grows, the critical component of talking-face videos continues to grow exponentially. Due to the advancement in internet services and camera technology, most of the videos today are captured and streamed in extremely high-resolution. Resolutions like 3840×2160 (4K) and 7680×4320 (8K) are considered to be mainstream and an important requirement for the entertainment industry.

With this growth in international video content, the ability to consistently dub the generated video content based on audio is a new multimedia application. Using this technology, video content can be watched seamlessly in other languages, as well as avatars can be anonymized for video conferencing, gaming, and other multimedia applications. However, the major challenge for audio-based visual dubbing has been the lack of scalability of audio-based lip-synchronization approaches. These either failed to generalize easily to multiple identities or, if suited for numerous identities, were unable to generalize to high-quality, high-resolution visual dubbing. Our work aims to solve this challenge comprehensively by enabling high-resolution lip-sync for any identity to a given speech. Before delving into the details, we start by surveying the major branches of current approaches for lip-syncing talking-faces to a given speech.

Speaker-Specific Lip-Sync Models Audio-driven talking-face generation has witnessed tremendous progress in recent years. The first works [12, 19] in this space dealt with large amounts of data of a specific speaker (e.g., President Obama) and trained deep neural networks to learn the speaker attributes. These works showed that learning phoneme-viseme correspondence through a neural network is possible. Follow-up works continued to deal with speaker-specific approaches [8, 13, 18, 20] with an aim to reduce the amount of speaker-specific data required for training. While the initial models were trained with over 20 hours of Obama's speeches, the latest approaches only need a few minutes of data per-speaker to generate high-quality results. The basic idea of all the speaker-specific approaches is to train two separate modules. The first module learns correspondence between lip shapes and speech, while a second renderer module generates the final video. Generally, this renderer is trained in a speaker-specific fashion. Although the data for isolated speakers has substantially reduced over the years, these models still fail to perform for unseen speakers, as well as for seen speakers with significantly altered appearances. Further, they also fail to handle dynamic environments like movie scenes consisting of large head motions and lighting variations.

Speaker-Agnostic Lip-Sync Models To learn the lip synchronization for in-the-wild speakers, speaker-agnostic works started gaining importance. These works [2, 11, 16] train on large datasets like LRS2 [3] containing thousands of identities to learn speaker-agnostic characteristics. They can handle unseen identities without requiring additional fine-tuning on speaker-specific data. They also work for various languages, poses and voices. The current state-of-the-art, Wav2Lip [16] is well known for generating lip-sync for videos of any identity in any language. Wav2Lip uses a standard encoder-decoder architecture that takes the target pose and target speech as input and generates a lip-synced face. A pre-trained lip-sync expert discriminator is used as a critique that penalizes the network for inaccurate lip shapes. However, Wav2Lip generates videos with a resolution of 96×96 pixels - making it practically unusable in professional videos that often require 4K resolution. We summarize the capabilities of the current models and compare them with our proposed method in Table 1.

Please note that we differ from audio-based talking Head generation works [24, 27, 30], where the aim is to generate the head movements along with lips from speech. Similarly, face re-enactment works [17, 23, 28] use a driving video to transfer the head motion to a source identity. In our case,

[1] https://www.business2community.com/content-marketing

[2] https://backlinko.com/zoom-users

Method	**Unseen IDs?**	**In-the-wild?**	**High Res.**
Synth. Obama [19]	×	×	✓
ObamaNet [12]	×	×	✓
Neural Puppetry [20]	×	×	✓
LipGAN [11]	✓	×	×
Wav2Lip [16]	✓	✓	×
Ours	✓	✓	✓

Table 1: Comparison of different lip-sync models. Our model handles the most challenging cases in this space.

we only morph lip movements to be in sync with a target speech without altering expressions or head motion, thus we exclude these works in our comparison.

Why not train Wav2Lip in ultra-high resolution? As Wav2Lip [16] is the current state-of-the-art in lip synchronization, the most straightforward and a natural question that arises is: "can we directly extend Wav2Lip to generate and lip-sync ultra-high resolution videos?" There are two major ways of achieving this: (i) Training Wav2Lip at higher resolutions (like 4K) and (ii) Using state-of-the super-resolution (SR) techniques on top of the current Wav2Lip generations. We observe that using either of these strategies results in sub-optimal generations. There are several key reasons to this. First, the lip-sync expert from Wav2Lip does not converge on high-resolution data from datasets like AVSpeech [5] or our proposed 4KTF dataset. We believe this is directly related to the increased number of pixels that the network deals with, increasing the overall variability. The encoder-decoder structure of Wav2Lip also faces similar issues and does not generate effective outputs. Another major challenge to deal with is the compute and hardware requirements. Training networks to generate videos at such high-resolutions runs into hardware issues. Also, such networks are extremely slow to train and work with small batch sizes, leading to poor performance.

As an alternative, using SR methods to upsample the Wav2Lip outputs is also not an ideal solution. The major reasons being: (i) Although Wav2Lip generates accurate lip and jaw regions, the resultant videos lack fine-grained facial features like teeth, lip color and face texture (in the generated lower-half of the face). These artifacts magnify when we apply the SR methods to obtain high-quality results; (ii) Wav2Lip generates videos at a resolution of 96×96 pixels. Upsampling these outputs to ultra-high resolutions like 4K would need video SR methods that can work at high scale-factors (like $8\times$ and $16\times$). However, the existing video SR methods [1, 9] are known to work effectively and generate high-quality results only at low scale-factors like $4\times$.

Our Contributions To address the problem of obtaining ultra-high resolution videos, we modify the existing approaches in the following way: We obtain a quantized generative pipeline that decodes *ultra-high resolution images*. The intermediate quantized representations in the generative pipeline are used to learn lip-synchronization using appropriate discriminators in the quantized latent space. Overall our generated faces contain 64-times more pixels than the current 96×96 output from Wav2Lip [16]. Our model works for any in-the-wild unseen identities, languages, and voices (including synthetic text-to-speech voice). Since the existing talking-face video datasets are limited in resolution, we collected a new 4K dataset from publicly available videos on YouTube. Our dataset spans a total of ≈ 30 hours, covering a diverse set of identities and an extensive vocabulary (see Figure 2. We train our model to synthesize high-quality talking-face videos with this dataset in hand.

2. 4K Talking Face Dataset

Previous datasets like MEAD [22], AVspeech [5] and HDTF [27] have done an incredible job collecting high-fidelity data but were limited in terms of resolution. We introduce the 4K Talking Face Dataset (4KTF), a new audio-visual dataset in 4K resolution. Our dataset consists of 140 YouTube high-quality (resolution: 4K) videos, amounting to $\approx$ 30 hours. The videos are of varying lengths, ranging from 40 seconds to 40 minutes, with over 2.5 million frames containing a taking face. The dataset predominantly contains English language videos and has a vocabulary of $\sim 10,000$ words. The videos are selected from different channels, including technical reviews, interviews, podcasts, educational content, and movie scenes. This results in a wide range of topics, a large vocabulary, and different speaking styles. Although most of the videos comprise a single speaker, we use active speaker detection [4] for the multi-speaker case to discard the segments in which the visible face and audio are out of sync. In addition, we use the YouTube transcripts to remove the segments containing inappropriate or violent language. We perform face detection using S3FD [26] to obtain the facial crops. At 4K resolution, face detection is not only slower but also surprisingly inaccurate. Therefore we resize the videos by a factor of 4 to perform face detection and then scale the coordinates back to the original resolution. We use the pre-processed videos with the face crops for the pipelines described in the next section. Please also note that the full resolution of the collected videos is at 4K while the face crops in the videos are of 768×768 pixel dimensions. Figure 2 shows different statistics from the dataset, along with some sample frames. We use this newly collected data to train all the networks. Since our dataset contains talking-face videos, speech, and automatically generated text transcripts (not used in this work), it will also be useful for several other related problems in the space involving the face, lip movements, speech, and text! We will release the dataset to aid future research in the audio-visual field.

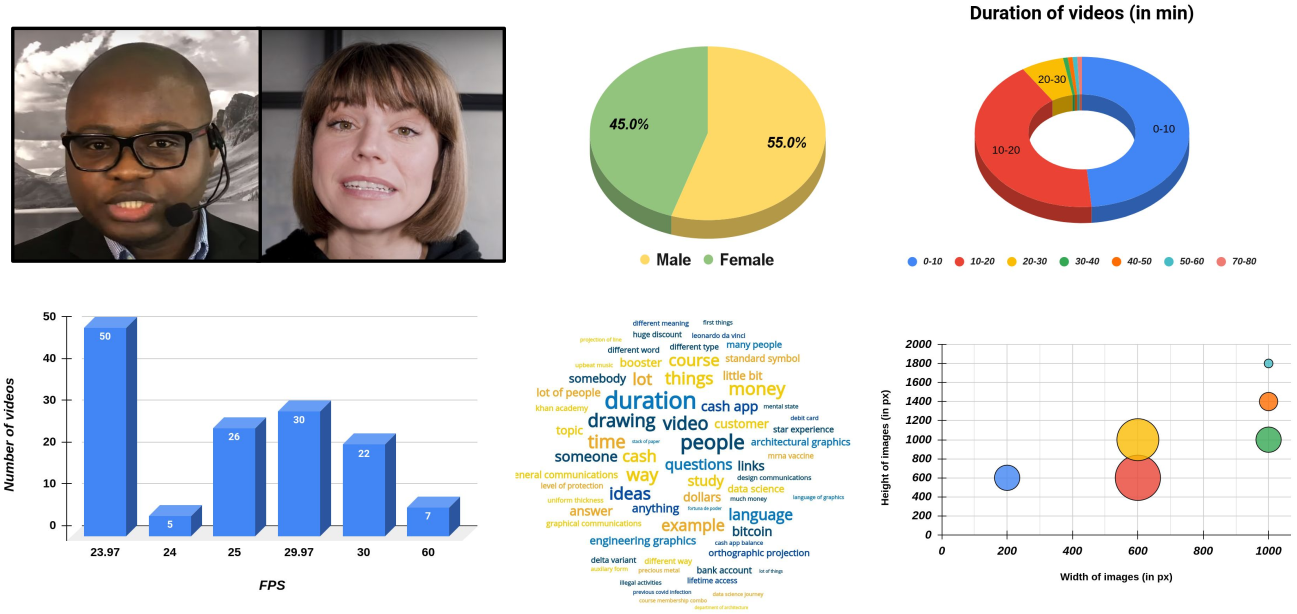

Figure 2: Samples and statistics of our newly collected 4K dataset (videos gathered from YouTube). Our dataset has a nearly equal male-female ratio, contains varying video lengths and FPS, spans an extensive vocabulary, and contains high-resolution frames. For more details about the dataset, please refer to our supplementary.

3. Generating Ultra-High Resolution Talking-Faces

Recent advances in high-resolution image synthesis have shown that learning compact vector spaces [6] helps in high-resolution synthesis. Methods like [6] first learn a VQ-GAN and then use it to generate intermediate quantized embeddings to represent the HD images. Downstream tasks like image-to-image translation, super-resolution, or random image generation are done using the quantized embeddings, i.e., in the quantized space. The final output from such downstream tasks is generated using the VQGAN decoder to convert resultant embeddings into RGB images.

3.1. Stage-1: Lip-sync Generator

Representing a Face and Head Pose in Quantized Space: In our work, we take a leaf out of this strategy and first learn a compact quantized space to represent higher resolution faces. We start by training a VQGAN [6], V_f, using the publicly available implementation[3]. The VQGAN encoder converts an input face image, $F_{in} \in \mathbb{R}^{H \times W \times 3}$ to an intermediate embedding $E_q \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times 256}$ through a set of convolution layers. A learnable codebook of $N_c \times 256$ is used to perform vector quantization on E_{in}. In our setup, we choose $H = W = 256$ and $N_c = 1024$ as the number of codebook entries. We obtain the vector quantized output E_q, which is then passed to a standard VQGAN decoder that reconstructs F_{in} (identical to [6]). Details regarding the losses and hyperparameters can be found in the same.

Similar to Wav2Lip [16] and LipGAN [11], we aim to morph the lip movements of the speaker and not change the target head pose. During training the lip-sync generator, it is paramount not to leak information regarding the mouth shape in the ground-truth face while providing the network with an accurate target head pose. Both WavLip and LipGAN achieve this by masking the lower half of the ground-truth face and conditioning the generator on the speech signal to generate it back. Unfortunately, we cannot directly use this trick in the quantized space. Masking the lower half of E_q does not stop the leakage of mouth information encoded in the top half of the embedding. Thus, we train a separate Pose-VQGAN, V_p, with only the top half of the face to avoid any unnecessary leakage. The encoder of V_p ingests a face image with lower half masked, $F_p \in \mathbb{R}^{H \times W \times 3}$ and outputs a quantized embedding $E_p \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times 256}$. The decoder then learns to generate the input F_p back from the quantized embedding. The network is trained with the losses mentioned in [6].

Once both V_f and V_p are trained to encode full faces and head poses, the next step is to train the lip-sync generator in the quantized space. We follow a similar training strategy as that of Wav2Lip [16]. We first train a lip-sync expert which acts as a critic in training the lip-sync generator.

Training a lip-sync expert in the quantized space: Our lip-sync expert uses a similar architecture proposed

[3] `https://github.com/CompVis/taming-transformers`

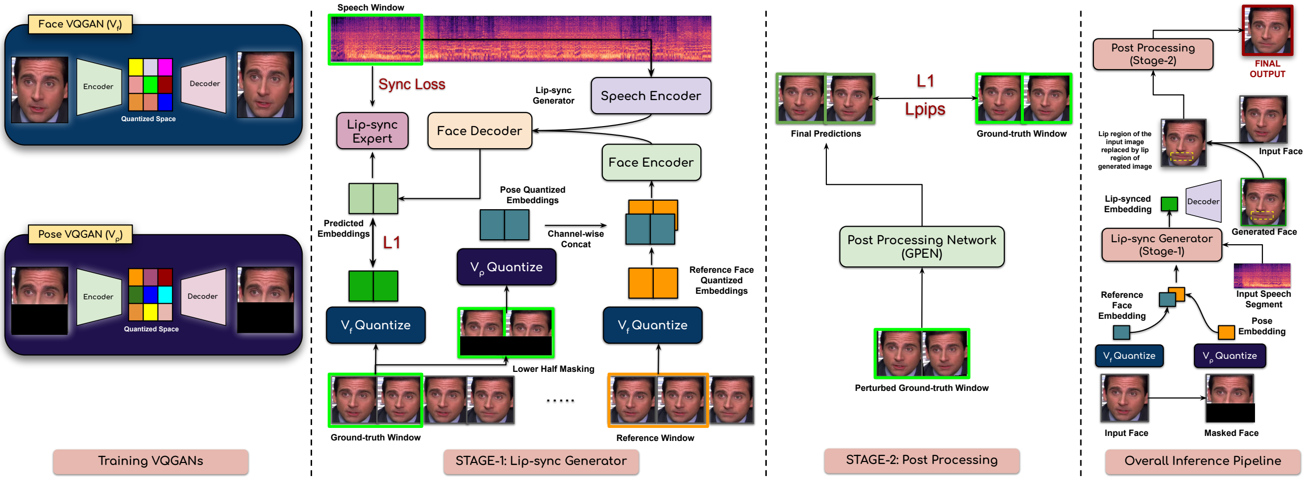

Figure 3: We present our pipeline for generating ultra-high resolution lip-synced videos. We first train Face VQGAN and Pose VQGAN networks (col-1) to encode the faces and head poses in a compact 16×16 dimensional space. We then train a lip-sync generator in the quantized space and get back the image using the Face VQGAN decoder. (stage-1, col-2). An optional post-processing network is used to improve the quality of the generated outputs (stage-2, col-3). To better understand our framework, we also show the overall inference pipeline (col-4).

in Wav2Lip [16], but we train our expert in the quantized space V_f, rather than the RGB space used in Wav2Lip. The network majorly contains video and speech encoders. The video encoder ingests the quantized embeddings of T_f consecutive frames and outputs a $D-$dimensional vector denoted by w_f. The speech encoder takes in a T_s length melspectrogram obtained from the input speech segment and generates a $D-$dimensional vector w_s. The final layer of both the encoders are ReLU activated, to ensure the vectors only have positive elements. For training the lip-sync expert, we sample video-speech pairs from the same time step (in-sync, i.e., positive pairs) and random pairs from different time-steps (out-of-sync, i.e., negative pairs). The network is trained using contrastive learning. We calculate the cosine similarity between w_s and w_f and back-propagate a binary cross-entropy loss to train the network. The lip-sync expert is trained on only 25 Frames-per-Second (FPS) videos with $T_f = 25$ frames and $T_s = 1$ second (100 melspectrogram time-steps).

Architecture of the Generator: Our generator network comprises three components: (i) face encoder, (ii) speech encoder, and (iii) face decoder. Both the face and speech encoders output $256-$dimensional embeddings. These are concatenated to form a $512-$dimensional encoding, which is given as input to the decoder. The encoders and decoder contain a stack of 2D convolution layers with residual blocks, batch normalization layers, and ReLU activation. In addition, we also add the skip connections between the face encoder and the face decoder for better gradient flow and to preserve the crucial facial features. The decoder finally generates a quantized embedding in the latent space V_f.

Training details: The generator network is trained to generate accurate lip shapes conditioned on a given speech segment. To prepare the input to the speech encoder, we take a short window of $T_x = 20$ melspectrogram time steps (200 ms of speech), denoted by S_x. We then take the middle frame, F_{gt}, of this speech window and consider it as the ground truth frame. We pass F_{gt} through the encoder of V_f to get the ground truth embedding E_{gt}, and mask the lower half of F_{gt}, pass it through V_p, and generate the pose embedding E_{gtp}. A reference frame, F_r, from a different time-step is selected and given to V_f, which generates the reference quantized embedding E_r. We channel-wise concatenate E_{gtp} and E_r, which acts as input to the face encoder. The speech encoder ingests the input speech melspectrogram S_x. We then concatenate the output of both the encoders. The decoder uses this concatenated embedding to predict the output embedding E'_{gt}. The network is trained using the $L1$ loss between E'_{gt} and E_{gt}. We also compute the sync loss using our pre-trained lip-sync expert discriminator which takes the audio-video pair (S_x, T_f) and detects if they are in-sync or out-of-sync.

Inference details: We consider a sliding window of 200ms (20 mel time-steps) across the full speech segment during inference. Each speech window is inferred separately through our lip-sync generator. Assuming we have a video during inference, we take the corresponding video frame and pass it to V_f, which generates the reference embedding. We also input a masked version of the frame to V_p, which encodes the pose. Both the reference and the pose embeddings are channel-wise concatenated and given to the lip-sync generator along with the melspectogram input. The decoder finally outputs a lip-synced quantized embedding.

3.2. Stage-2 (Optional): Post-Processing stage 1 output

This is an optional stage to further improve the visual quality of the generated output from the stage 1. We use GPEN [25] as the post processing network and found that we get slightly improved and sharper results. We train GPEN [25] following the original training procedure and losses on our newly collected 4KTF dataset. During inference, we feed a modified face crops to the network: we replace only the lip region of the original face crops of the video with the output generated from stage 1 using the lip landmarks obtained from Mediapipe [14]. The synthesized outputs are then pasted back into the original video. More details about the architecture, training and inference procedures can be found in the supplementary material. This stage is totally optional and can be replaced with any post-processing network.

3.3. Watermarking the Final Outputs

Talking-face generation models [10, 16, 17, 23, 29] enable a plethora of positive applications. However, there are potential negative impacts due to the possibility of harmful "deepfakes". We add an invisible watermark to our dataset using the invisible watermarking technique [15] [4]. There is no change in the perceptual visual quality of the image. A randomly generated fixed string is embedded (watermarked) into the image in the frequency space using the DWT + DCT + SVD transformations. We can decode the image to get back the fixed string using the inverse of each transform. We first watermark the whole dataset and then train the network. It ensures the watermark is inherently learned by the model and outputs it in each of the generated face crops, which are finally pasted back into the full frame. While testing, we first detect each face region present in a video. We then try to decode the watermark in the detected facial areas in each frame. If 50% of the total frames contain the watermark, we assume it to be a match.

4. Experiments

In this section, we evaluate various aspects of the generated outputs from our method on different datasets. We also include several visual results from our technique and compare them with the current state-of-the-art methods.

4.1. Quantitative Evaluations

Metrics: To evaluate the quality of lip-synchronization, we use "Lip Sync Error - Confidence" (LSE-C) and "Lip Sync Error - Distance" (LSE-D) metrics introduced in Wav2Lip [16]. The publicly available pre-trained model SyncNet [4] is used to calculate the lip-sync errors. More details about the two metrics can be found in Wav2Lip [16]. In addition to these metrics, we also use the popular Fréchet Inception Distance (FID) to evaluate the perceptual quality of the generations at a frame level. Similarly, we use the Fréchet Video Distance (FVD) [21] proposed to measure the perceptual quality at the video level. FVD is used to measure both temporal coherence as well as sharpness at the frame level. These metrics are calculated using only the face crops, ensuring the high-resolution background does not play any role in the calculations.

Baselines: We compare our work with multiple baselines. We modify the publicly available codebases of "You said that?" [2], "LipGAN" [11] and "Wav2Lip" [16] and train them using the same settings and datasets as our model (768 × 768 pixel resolution). We make suitable changes in the architectures to handle the higher resolution input. As another baseline, we use the publicly available Wav2Lip model at original resolution (96 × 96) and use the pre-trained state-of-the-art video super-resolution model "TecoGAN" [1] to obtain the super-resolved videos at the target resolution. We evaluate all the models on 5000 selected videos from the AVSpeech test-set and the test-set from the proposed 4K dataset. Please note that the AVSpeech test set is evaluated at 1080p resolution.

Results: As seen in Table 2, we outperform the competing methods by a significant margin. Our method produces lip-synced videos at very high-resolutions (indicated by LSE metrics). The generated outputs are sharper and highly temporally coherent compared to the previous works (indicated by FID and FVD metrics). Our method surpasses the existing baselines in generating high-quality frames with very few artefacts (also validated in Figure 4 and supplementary video).

Performance on Silent Regions: While Wav2Lip [16] generates accurate lip-sync in most cases, it struggles with long silent regions. The original lip movements present in the video interferes with the generated ones resulting in significant quivering of lips. We provide silent audio as input to all the videos in the test set and compare our results to that of Wav2Lip in Table 3. A visual demonstration of samples is also provided in Figure 5. As seen from both the table and the figure, our model handles silences far better than Wav2Lip. We hypothesize the reason for this to be learning lip-sync in the quantized space, which is richer than the image space that Wav2Lip was trained on.

4.2. Human Evaluations

Since the quality of lip-sync is highly subjective, we perform human evaluations on the generated videos. We show the outputs from different algorithms to 50 users and ask them to rate the videos on a scale of $1-5$, with 1 being the lowest rating and 5 being the highest. The users are asked to rate the following three attributes: (i) Lip sync Quality, (ii)

[4] `https://github.com/ShieldMnt/invisible-watermark`

	AVSpeech [5]							4KTF						
Method	**LSE-C ↑**	**LSE-D ↓**	**FID ↓**	**FVD ↓**	**LSQ ↑**	**Shrp. ↑**	**OE ↑**	**LSE-C ↑**	**LSE-D ↓**	**FID ↓**	**FVD ↓**	**LSQ ↑**	**Shrp. ↑**	**OE ↑**
You-said-that-4K [2]	0.98	10.01	9.12	9.81	2.50	1.32	1.98	1.07	10.47	18.34	9.83	1.32	1.44	1.41
LipGAN-4K [11]	1.09	9.52	7.63	8.52	2.63	1.71	2.31	1.43	8.18	14.21	9.16	1.47	1.42	1.31
Wav2Lip-4K [16]	2.66	9.13	8.01	8.41	3.17	1.65	2.18	3.12	8.74	7.54	7.91	3.52	1.37	2.63
Wav2Lip-orig [16] + TecoGAN [1]	4.17	6.33	7.47	7.16	3.26	1.94	2.27	4.03	7.24	7.18	8.86	3.43	1.72	2.14
Ours	**7.26**	**6.21**	**5.18**	**6.41**	**3.72**	**4.51**	**4.32**	**7.10**	**6.32**	**6.84**	**6.66**	**6.86**	**4.43**	**4.62**

Table 2: Quantitative comparison of different methods on AVSpeech [5] and our new 4KTF datasets. Our model outperforms all baselines by a large margin. Using our approach, we can obtain high-quality outputs (indicated by FID and FVD) and accurate lip synchronization (indicated by LSE-C and LSE-D). Note that FVD is scaled by a factor of 100 for better readability. We also report the human evaluation scores based on: (i) Lip-sync Quality (LSQ), (ii) Sharpness (Shrp.), and (iii) Overall Experience (OE).

Sharpness and other details of the face, and (iii) Overall Experience of the video. We report the mean opinion scores in Table 2. In line with the quantitative evaluation, our method achieves the highest scores in all these attributes, indicating the robustness of our approach.

Figure 4 depicts the samples generated from different models. We can observe that our model generates a highly detailed lip region compared to the current methods. It effectively reconstructs fine-grained facial features like teeth, lip color, lip and jaw texture, and has minimal to no artifacts. We find the visual results to corroborate the findings in our quantitative and human evaluations.

Method	**LSE-C ↑**	**LSE-D ↓**	**FID ↓**	**FVD ↓**
Wav2Lip-orig [16] + TecoGAN [1]	1.08	12.73	7.124	10.88
Ours	**4.18**	**8.21**	**6.79**	**9.03**

Table 3: Our method works well on silent regions of the video.

5. Ablation Studies

We perform several ablations to verify the effect of our different components. The scores are reported on the test set of 4KTF dataset.

Importance of Post Processing Network To assess the importance of the Stage 2 network, we compare the results of stage 1 and 2 of our pipeline. While the results have decent lip-sync, the stage 2 results are slightly sharper, as also can be seen in Table 4.

Method	**LSE-C ↑**	**LSE-D ↓**	**FID ↓**	**FVD ↓**
Ours w/o Stage 2	7.01	6.31	7.12	7.48
Ours	**7.10**	**6.32**	**6.84**	**6.66**

Table 4: Comparison of stage 1 and 2 results.

Importance of the lip-sync expert We train a lip-sync generator without using the sync loss and report results in Table 5. We also vary the context window size T - we test with $T = 5$ and $T = 25$. We find that the lip-sync expert trained on longer audio-visual sequences perform better and is selected for the final version. We also calculate the accuracy of the lip-sync expert by creating random audio-visual pairs that are in-sync and out-of-sync with 50% probability. Table 5 indicates that the best accuracy is achieved for the model trained with sync loss using a context window of 25 frames.

Method	**LSE-C ↑**	**LSE-D ↓**	**Acc. ↑**
Ours w/o Sync Loss	1.13	11.01	-
Ours with Sync Loss, T=5	3.12	10.38	65.1%
Ours with Sync Loss, T=25	**7.10**	**6.32**	**91.2%**

Table 5: We evaluate the importance of lip-sync expert and also show the effect of using different context windows.

We find that, a lip-sync expert which is trained with 25 frames is the most accurate forces the generator to produce most accurate lip shapes.

6. Applications

We believe our model is a perfect fit for several applications at a time when the amount of multimedia content around the globe is growing exponentially. Few of the potential applications enabled by our model are as follows. (i) **Movie and television industries:** Modern movies are dubbed and released in tens of languages. Our model can lip-sync such dubbed movies with ease and improve the viewing experience. Similarly, other forms of dubbed content like TV shows, interviews, documentaries, and lectures can also be precisely lip-synced; (ii) **Marketing:** Marketing videos are essential for reaching out to customers. Generating realistic marketing videos at scale can reduce the cost and is sought after by businesses all around the globe. Instead of recording hundreds of marketing videos for different products, a single video can be lip-synced with various audios and languages, thus reducing the cost; (iii) **Online meetings:** Hours of online meetings have given rise to issues like Zoom Fatigue [7], i.e., getting tired of looking into the camera. Our work can potentially be used to replace the actual video stream of the speaker with a gen-

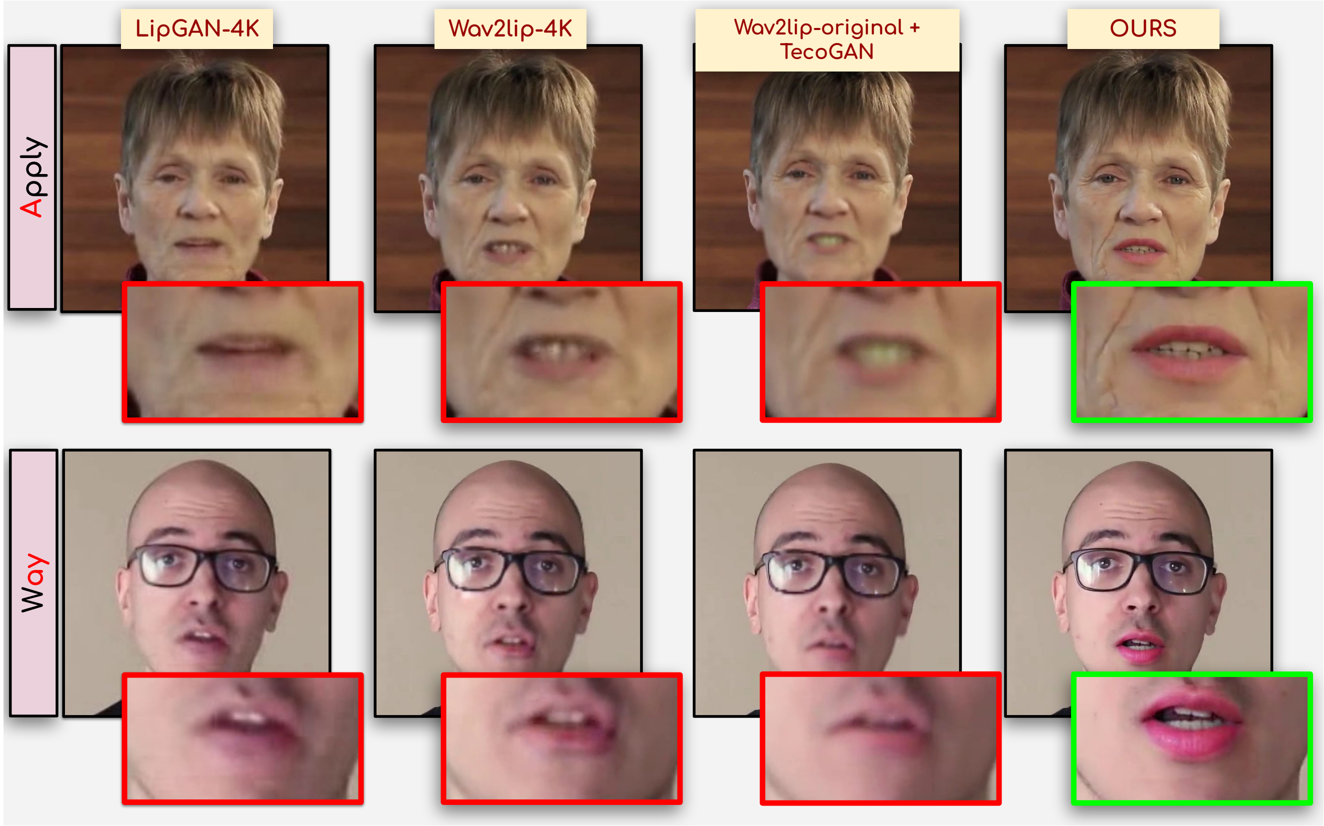

Figure 4: Sample results from different algorithms. Clearly, our model generates far better, sharper, and higher-quality outputs. Our model captures intricate details like teeth, wrinkles of skin and lip color, which the previous models fail to generate.

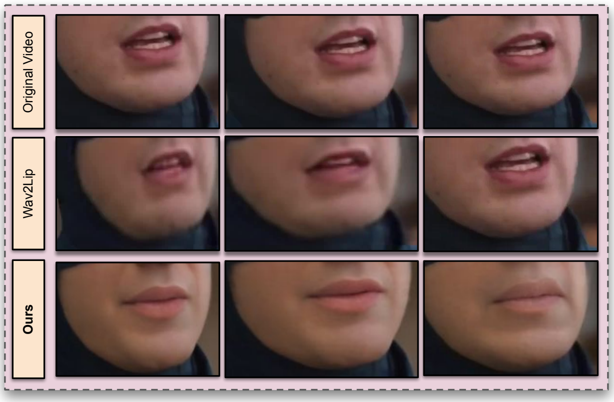

Figure 5: Performance evaluation on silent speech segments. While the output from Wav2Lip follows the original lip movements, our model can generate closed lip shapes in sync with the silent speech.

erated content that is in sync with the spoken content. Our model can also generate the video stream in case of a drop in connection quality; (iv) **Animations:** Even though our model is never trained on CGI faces, it still performs well on animated characters. This allows our model to be used in gaming and animated movies; and (v) **Training:** Since our model generates accurate lip shapes given a speech segment, it can be used to teach lip-reading to people hard of hearing and their family members. A wide variety of course content showing the lip movements corresponding to words and sentences can be created, enabling large-scale training of human lip readers.

7. Conclusion

This work presents the first approach in generating ultra-high resolution talking-face videos. With our approach, it is now possible to synthesize talking-face videos with accurate lip shapes at very high-resolutions (4K). Our work revolves around a two-stage framework where we first learn to lip-sync in a compact vectorized space and then render the high-resolution face outputs. We generate state-of-the-art, realistic, high-quality results at such high resolutions for the first time and mark significant improvements over the competitive methods. We believe our work will positively impact several industries, open up new applications and make movie-making much easier!

References

[1] Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for gan-based video generation. *ACM Trans. Graph.*, 39, July 2020.

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

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

[4] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016.

[5] 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. 37(4), July 2018.

[6] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020.

[7] G. Fauville, M. Luo, A.C.M. Queiroz, J.N. Bailenson, and J. Hancock. Zoom exhaustion & fatigue scale. *Computers in Human Behavior Reports*, 4:100119, 2021.

[8] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. Text-based editing of talking-head video. *ACM Trans. Graph.*, 38(4):68:1–68:14, July 2019.

[9] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[10] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. *arXiv preprint arXiv:2104.07452*, 2021.

[11] Prajwal K R, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and C V Jawahar. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19. ACM, 2019.

[12] Rithesh Kumar, J. Sotelo, K. Kumar, A. D. Brébisson, and Yoshua Bengio. Obamanet: Photo-realistic lip-sync from text. *ArXiv*, abs/1801.01442, 2018.

[13] Avisek Lahiri, Vivek Kwatra, Christian Früh, John Lewis, and Chris Bregler. Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. *CoRR*, abs/2106.04185, 2021.

[14] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines. *ArXiv*, abs/1906.08172, 2019.

[15] KA Navas, Mathews Cheriyan Ajay, M Lekshmi, Tampy S Archana, and M Sasikumar. Dwt-dct-svd based watermarking. In *2008 3rd International Conference on Communication Systems Software and Middleware and Workshops (COMSWARE'08)*, pages 271–274. IEEE, 2008.

[16] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 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, page 484–492, 2020.

[17] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, December 2019.

[18] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *arXiv preprint*, arXiv:, 2020.

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

[20] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. *arXiv preprint arXiv:1912.05566*, 2019.

[21] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[22] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, August 2020.

[23] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[24] Haozhe Wu, Jia Jia, Haoyu Wang, Yishun Dou, Chao Duan, and Qingshan Deng. *Imitating Arbitrary Talking Style for Realistic Audio-Driven Talking Face Synthesis*, page 1478–1486. Association for Computing Machinery, New York, NY, USA, 2021.

[25] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 672–681, 2021.

[26] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and S. Li. S^{3}fd: Single shot scale-invariant face detector. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 192–201, 2017.

[27] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021.

[28] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[29] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[30] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makeittalk: Speaker-aware talking-head animation. *ACM Transactions on Graphics*, 39(6), 2020.