Audio-Visual Face Reenactment

Madhav Agarwal
IIIT, Hyderabad

Rudrabha Mukhopadhyay
IIIT, Hyderabad

Vinay Namboodiri
University of Bath

C V Jawahar
IIIT, Hyderabad

{madhav.agarwal,radrabha.m}@research.iiit.ac.in, vpn22@bath.ac.uk, jawahar@iiit.ac.in

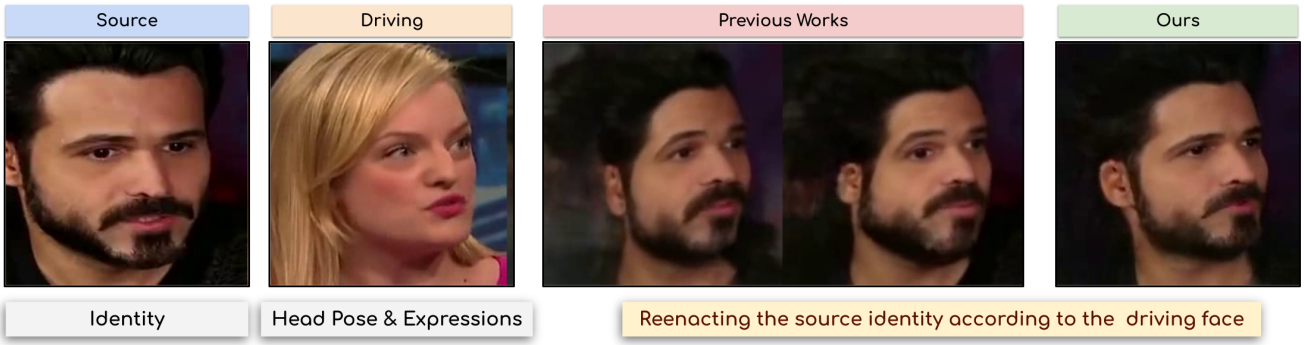

Figure 1: We propose AVFR-GAN, a novel method for face reenactment. Our network takes a source identity, a driving frame, and a small audio chunk associated with the driving frame to animate the source identity according to the driving frame. Our network generates highly realistic outputs compared to previous works like [29] and [30]. Results from our network contain significantly fewer artifacts and handle things like mouth movements, eye movements, etc. in a better manner.

Abstract

This work proposes a novel method to generate realistic talking head videos using audio and visual streams. We animate a source image by transferring head motion from a driving video using a dense motion field generated using learnable keypoints. We improve the quality of lip sync using audio as an additional input, helping the network to attend to the mouth region. We use additional priors using face segmentation and face mesh to improve the structure of the reconstructed faces. Finally, we improve the visual quality of the generations by incorporating a carefully designed identity-aware generator module. The identity-aware generator takes the source image and the warped motion features as input to generate a high-quality output with fine-grained details. Our method produces state-of-the-art results and generalizes well to unseen faces, languages, and voices. We comprehensively evaluate our approach using multiple metrics and outperforming the current techniques both qualitative and quantitatively. Our work opens up several applications, including enabling low bandwidth video calls. We release a demo video and additional information at `http://cvit.iiit.ac.in/research/projects/cvit-projects/avfr`.

1. Introduction

Imagine your favorite celebrity giving daily news updates, motivating you to work out, or interacting with you on your mobile phone! What if a movie director could reenact an actor's image without actually recording the actor? Or, how about skilled content creators animating avatars in a metaverse to follow an actor's head movements and expressions in great detail? We can also reduce zoom fatigue [11] by animating a well-dressed image of ourselves in a video call without transmitting a live video stream! These ideas seem fictitious, infeasible, and not scalable. But, how about animating or "reenacting" a single image of any person according to a driving video of someone else? Face reenactment, thus, opens up many opportunities in a world that is becoming increasingly digital with each passing day.

Face Reenactment aims to animate a source image using a driving video's motion while preserving the source identity. Multiple publications have improved the quality of the generations. Existing works on talking head generation generally use a single modality, i.e., either visual[12, 29, 39, 40] or audio features[13, 37, 31]. Audio-driven talking

head generation models are good at generating quality lip-sync; however, they have a serious drawback in handling non-verbal cues. The video-driven methods heavily rely on the disentanglement of motion from the appearance [17]. These methods generally use key points as an intermediate representation [29, 12, 39] and try to align the detected key points of source and driving frames. These works learn key points in an unsupervised manner and fail to focus on specific regions of the face. This stems from inadequate priors regarding the face structure or the uttered speech. The final quality of the generations also suffers from using a basic CNN-based decoder that fails to capture the sharpness present in the source image and generates blurred output video. As a part of this work, we provide a detailed review of different approaches in Section 2.

In this paper, we analyze the shortcomings of the current works and add key modules to our network. We introduce **A**udio-**V**isual **F**ace **R**eenactment **GAN** (**AVFR-GAN**), a novel architecture that uses both audio and visual cues to generate highly realistic face reenactments. We start with providing additional priors about the structure of the face in the form of a face segmentation mask and face mesh. We also provide corresponding speech to our algorithm to help it attend to the mouth region and improve lip synchronization. Finally, our pipeline uses a novel identity-aware face generator to improve the final outputs. Our approach generates superior results compared to the current state-of-the-art works, as shown in Section 4. We comprehensively evaluate our method against several baselines and report the quantitative performance based on multiple standard metrics. We also perform human evaluations to evaluate qualitative results in the same section. Our proposed method opens a host of applications, as discussed in Section 6, including one in compressing video calls. Our work achieves more than $7\times$ improvement in visual quality when tested at the same compression levels using the recently released H.266 [7] codec.

Our contributions are summarized as follows:

1. We use additional priors in the form of face mesh and face segmentation mask to preserve the geometry of the face.
2. We utilize additional input in the form of audio to improve the generation quality of the mouth region. Audio also helps to preserve lip synchronization, enhancing the viewing experience.
3. We build a novel carefully-designed identity-aware face generator to generate high-quality talking head videos in contrast to the high levels of blur present in the previous works.

2. Related Work

Talking head generation works can be broadly classified in three categories based on the type of input they use to generate a talking head: Text-driven [16, 33, 36], Audio-driven [9, 13, 18, 31, 37, 43, 45], and Video-driven [12, 27, 29, 39, 44] Talking Head Generation.

Text-driven Talking-head Generation Text-driven natural image generation [25, 26] has recently seen a lot of progress in the computer vision community. Inspired by the recent success of GANs in generating static faces from text[38], Li *et al*. [16] proposed a method to use text for driving animation parameters of the mouth, upper face and head. Txt2Vid [33] converts the spoken language and facial webcam data into text and transmits it to achieve low-bandwidth video conferencing using talking head generation. However, this method relies heavily on the generated speech, altering the original speaker's voice, prosody, and head movements in the video call. It depends on the quality of the Speech-to-Text module, which introduces grammatical errors and language dependency. Text as a medium has very little information about the head and lip movements; thus, we consider the problem ill-posed.

Audio-driven Talking-head Generation While text-driven methods suffer from a significant lack of adequate priors, we now move on to audio, a much more expressive and informative form of input. As the name suggests, audio-driven methods [9, 13, 18, 31, 37, 43, 45] use only audio to animate a static face image. The first set of works like You-said-that? [9], LipGAN [15] and Wav2Lip [24] achieved lip synchronization with given audio but failed to generate head movements in sync with the speech. These works used fully convolutional architectures and generated a single frame at a time without considering the temporal constraints. Eventually, a different class of works starting from Song *et al*. [31] in 2018 and Zhou *et al*. [43] in 2019, started using conditional Recurrent Neural Networks to model the temporal characteristics of a talking face. In 2020, Zhou *et al*. [45] published a landmark work that predicted dense flow from audio instead of directly generating the output video. The dense flow was then used to warp the source image to generate the final output. Several other well-known works like Emotional Video Portraits [13] add an additional emotion label as input to create the talking head in the desired emotion. However, all of these works lack fine-grained control of the talking head and often contain a loopy head motion, and thus cannot be directly used in many applications.

Video-driven Talking-head Generation Finally, we move to video-driven methods, which use a driving video to get the motion and other facial features required to reenact a source image. Please note that the driving video and the source image may not have the same identity. Owing to the significant priors in driving video, the final generation quality of video-driven methods surpasses those of

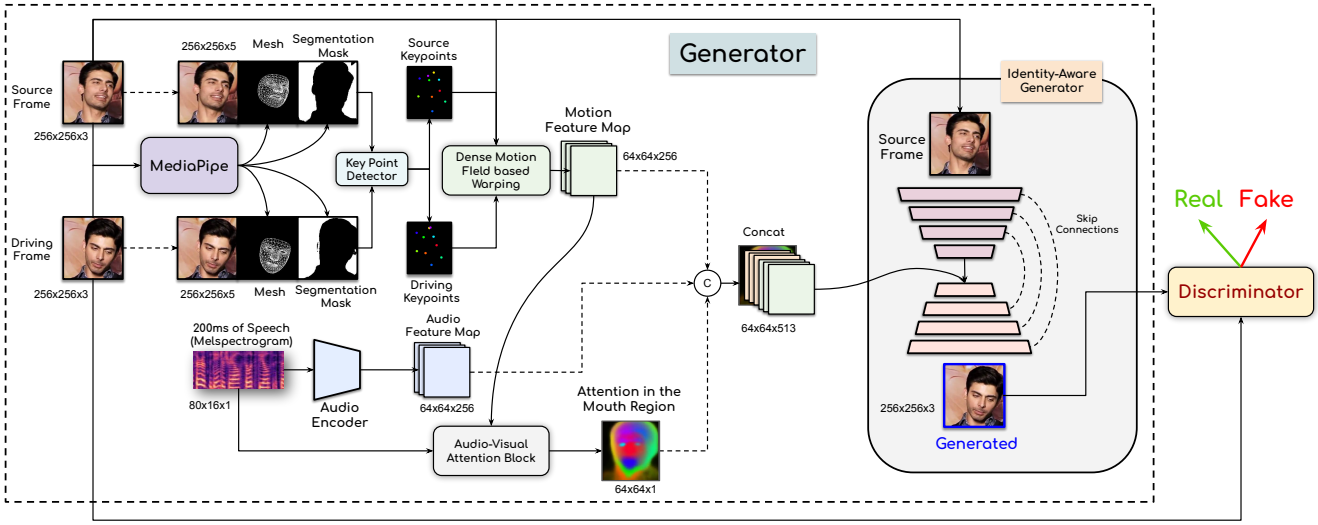

Figure 2: The overall pipeline of our proposed Audio Visual Face Reenactment network (AVFR-GAN) is given in this Figure. We take the source and driving images, along with their face mesh and segmentation masks to extract keypoints. An audio encoder extracts features from driving audio and use them provide attention on lip region. The audio and visual feature maps are warped together and passed to the carefully designed Identity-Aware Generator along with extracted features of the source image to generate the final output.

text-only and audio-only ones. The most influential work in this area, First-Order-Motion-Model (FOMM), was published by Siarohin *et al*. [29] in 2019. The key idea was to estimate the motion field from sparse keypoints detected in both source and driving frames. The motion field was used to calculate dense flow and warp the source frame in a latent space. Several other works [39, 12] followed the same principle and added supplementary components to improve the quality. Face-vid2vid [39] used keypoint information in a 3D space, taking care of head rotation, among other things. DA-GAN[12] further added depth-aware attention to provide dense 3D facial geometry to guide the generation of motion fields. A similar approach in Motion-Representation-in-Articulated-Animation [30] uses key regions instead of keypoints to generate the warpable motion field. Approaches like ICface[34] provide a method to control the pose and expressions of a face image using head pose angles and action unit values. Recently, Zhang *et al*. [42] proposed using the three-dimensional morphable face model (3DMM) parameters to reenact a face image. They demonstrated that motion descriptor parameters for 3DMM can be derived from a driving video and, in turn, animate a static facial image.

To the best of our knowledge, PC-AVS [44] is the only work that uses audio and video to formulate a low-dimension pose and motion code. Unlike FOMM, PC-AVS does not predict motion fields to calculate dense flow and warp the source image. Instead, they try to train their network to learn motion in a latent space inherently. While this allows them to achieve state-of-the-art lip sync, the generated video's overall quality is considered inferior to works like DA-GAN [12]. In this work, we base our approach on FOMM's [29] principles and improve it with additional audio information. We also provide additional structural information to extract better geometries of the face. This allows us to use the best of both worlds and propose a novel network AVFR-GAN as described in the next section.

3. Audio-Visual Face Reenactment GAN

We present **A**udio-**V**isual **F**ace **R**eenactment GAN (**AVFR-Gan**), which takes a source image and a driving video plus audio to create high-quality talking head videos by preserving the source identity. As mentioned previously, we follow a similar strategy to that of FOMM [29] for our training pipeline. Instead of generating multiple frames in the form of a video, we handle the input in a frame-by-frame fashion. Our main goal is to estimate the motion between a source and a driving frame and then warp the source frame accordingly to generate an approximation of the driving frame. Our model can be broadly divided into a Generator M_{Gen} and a discriminator M_{Disc} as shown in Figure 2. We first discuss the individual components present inside the generator.

Additional Structural Priors to the Keypoint Detector We start with selecting a source frame F_s and a driving frame F_d both of dimensions $h \times w$. During training, both of these frames are selected from the same video. We pass these frames through mediapipe [19] to generate a face

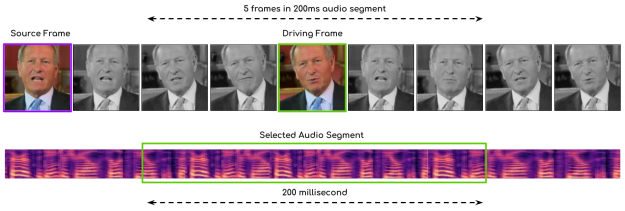

Figure 3: Illustration of Audio window selector mechanism. It generates a 200ms spectogram such that the driving frame remains in the middle of the segment. In case of a 25 FPS video, a 200ms segment contains 5 frames.

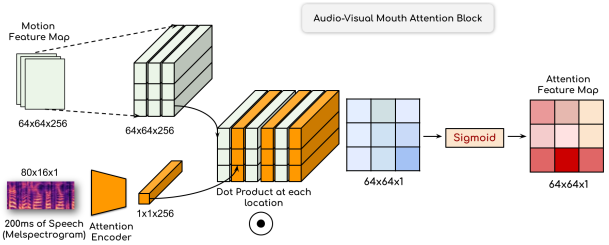

Figure 4: Illustration of Audio Visual Attention module. Attention is generated by taking the dot product between a learned audio feature and visual features in each location, followed by a Sigmoid activation.

mesh and a face segmentation map. We channel-wise concatenate the generated mesh and the segmentation mask with their respective images and create 5 channel versions of the same. We term the concatenated source and driving frames as I_s and I_d, respectively. We use these concatenated inputs to feed into our keypoint detector, M_{kp}. The addition of these priors helps us in providing the keypoint detector with more information about the respective structures of source and driving frames. Furthermore, the segmentation mask also provides the module with foreground and background information enabling the keypoints to be detected only from the foreground. We use the keypoint detector from FOMM [29] in our architecture. The keypoint detector M_{kp} detects K keypoints. More concretely, we can write,

$$\{X_{T,n}\}_{n=1}^{K} = M_{kp}(I_T), T \in s, d \tag{1}$$

The difference between the generated keypoints from the source and driving frames is used to calculate the motion field following FOMM. The motion field is then used to calculate dense flow and generate a warped feature map. We denote this feature map as Motion Feature Map, Enc_{motion} as it captures the motion between the source and the driving frames. The dimension of this feature map is kept to be $\frac{h}{4} \times \frac{w}{4} \times c$. We plot sample keypoints detected in specific frames in Figure 5 (left). Also, note that each keypoint has a specific region of interest in the generated motion field. We plot the heatmaps for each keypoint in Figure 5 (middle). The heatmaps show that the regions of interest for each keypoints correspond to specific facial features. For example, the dark blue keypoint attends to the mouth region, green attends to the jaw, and sky blue attends specifically to the eye regions. Interestingly both of the eyes are attended by the same keypoint.

Audio-conditioned Features Audio (mainly speech in our case) is an essential source of information that often accompanies a talking-head video. We decided to use the speech from the driving video to improve the quality of mouth movements in the generated video. While works like MakeItTalk [45] have already generated head movements solely from audio, our goal is to only improve the mouth movements and transfer head motion directly from the driving video. Therefore, we follow the same strategy taken by lip-synchronization works like [9, 15, 24] to handle speech. We select the 200ms window of speech around our driving frame F_d such that F_d is the middle frame in the sampling window. A graphical representation of the audio window selection is given in Figure 3. We generate melspectrogram I_{mel} from the speech window and feed it to a 2D CNN-based encoder. The audio encoder also outputs a feature map, Enc_{aud}, of $\frac{h}{4} \times \frac{w}{4} \times c$ dimension. We concatenate $(Enc_{motion}, Enc_{aud})$ along with the attention map generated as described next.

Audio-Visual Attention Apart from improving the lip synchronization in the generated video, we propose using audio to specifically attend to the speaker's mouth region, enhancing the fine-grained details like teeth in the generated video. To do this, we pass I_{mel} through an attention encoder generating an encoding Enc_{query} of dimensions $1 \times 1 \times c$. We then take Enc_{motion} of dimension $\frac{h}{4} \times \frac{w}{4} \times c$ and calculate the dot product at each location with Enc_{query}, generating a $\frac{h}{4} \times \frac{w}{4} \times 1$ matrix. We pass this through a Sigmoid layer to get the attention map Enc_{attn} as shown in Figure 4. A formal definition of this block is given in Equation 2.

$$Enc_{attn}(i,j) = Sigmoid(Enc_{query} \odot Enc_{motion}(i,j)), \\ i \in \frac{w}{4}, j \in \frac{h}{4} \tag{2}$$

A visualization of the audio-visual attention can be found in Figure 5. As we can see, audio not only helps the model to attend to the mouth region but also helps the network attend to other regions like the eyes, which correlates to expressions from speech.

Identity-Aware Generator We propose a novel generator to decode the concatenated feature vector. We analyze the current decoders used in FOMM [29], Face-

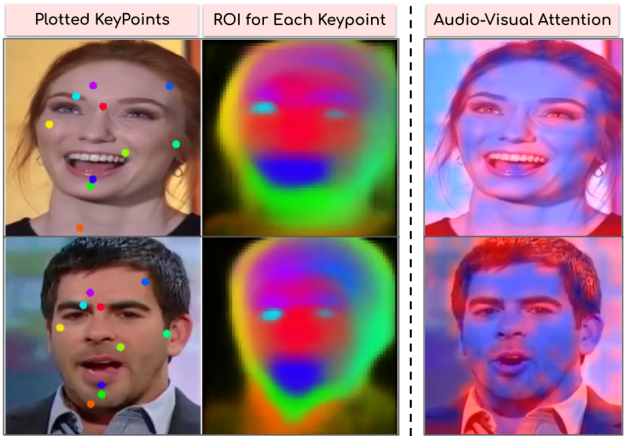

Figure 5: Illustration of keypoints detected (left), colour coded heatmap corresponding to each keypoint (centre) and the attention generated by our Audio-Visual Module (right). The ROI image shows that there are keypoints specific to the eye and mouth region. Attention image shows the important facial regions on which AVFR-Gan focuses.

Vid2Vid [39] and DA-GAN [12]. We realize that the pipelines followed by the current works fail to capture information from the source image directly. The network entirely depends on the warped features generated from the motion estimator to get the identity characteristics of the source speaker. Unfortunately, the warped features are forced to encode motion and fine-grained identity information, making it tougher to train. This ultimately causes the outputs to contain major artifacts and lose sharpness. We improve upon this and design an identity-aware face generator. We first concatenate Enc_{motion}, Enc_{con} and Enc_{attn} together to get the final warped features, generating Enc_{dec}. Instead of only feeding the warped features, we also feed in the source image F_s separately to the UNet-shaped [28] generator. The generator consists of an identity-encoder and a decoder. Both the encoder and decoder contain residual convolutional blocks inspired from Spatially Adaptive Normalization [23]. The source image F_s is first passed through an identity encoder to encode identity information. The output from the identity encoder is then concatenated with Enc_{dec} and finally passed through the matching decoder with appropriate skip connections between the encoder and decoder blocks. The final output from the generator is denoted by F_{gen}. Our generator produces the sharpest output compared to the current state-of-the-art, as shown in the subsequent sections.

Discriminator To improve the quality of our generated outputs, we also employ a standard discriminator, which is trained in a GAN setup along with the rest of the network. Our discriminator M_{Disc}, consists of a stack of Conv2D layers each followed by either spectral normalization [21] or instance normalization [35]. Each convolution block is followed by a Leaky ReLU activation [20]. The discriminator predicts a real or fake label and is trained to maximize the following loss function L_{Disc} given in Equation 3.

$$\max_{M_{Disc}} L_{Disc} = \mathbb{E}_{x \sim p_{real}} \log M_{Disc}(x) + \mathbb{E}_{F_{gen}} \log(1 - M_{Disc}(F_{gen})) \tag{3}$$

Losses used to train the Generator We use multiple loss functions similar to [29]. We use the $L1$ reconstruction loss between F_d and F_{gen}. We also use the LPIPs [41] perceptual similarity loss (denoted by L_{per}) to improve the perceptual quality of the generated outputs. Finally, we employ the equivarience constraints L_{eq} as described in the original FOMM paper. We refer the reader to [29] for information regarding these constraints. While training the generator we also minimize the discriminator loss given in Equation 3. Therefore, we present our final loss function, Equation 4.

$$\min_{M_{Gen}} L_{Gen} = ||F_d - F_{gen}||_1 + L_{per} + L_{eq} + \mathbb{E}_{F_{gen}} \log(1 - M_{Disc}(F_{gen})) \tag{4}$$

Inference Setting While we sample both F_s and F_d from the same video during training, our training strategy ensures that identity and motion information are well distilled. Therefore, our method allows for cross-identity face reenactment. During inference, we select a single image of a person as the source image F_s. Given a driving video of N frames, $V_{i...N}$, we pass each frame separately through our network along with F_s and the corresponding audio segment of V_i (denoted by A_i) to generate the final output as shown in Equation 5.

$$F^i_{Gen} = M_{Gen}(F_s, V_i, A_i), i \in 1...N \tag{5}$$

Implementation Details In our experiments, we set $h = 256, w = 256$ and predict $K = 10$ keypoints for training all our models. The model is trained using the Adam optimizer[14] with a learning rate scheduler set at 60 and 90 epochs. The initial learning rate is set to be at 0.001. The training time taken by model on 4 NVIDIA RTX 3080Ti GPUs with a batch size of 10 is around 10 days. We train our model on the VoxCeleb [22] dataset, which contains 25 FPS videos. Thus, the 200ms audio window consists of 5 frames, of which the 3rd frame is selected as the driving frame F_d. Any other random frame from the same video is selected as F_s during training the network. More details about the network structure and other training characteristics are provided in the supplementary material.

	Same-id Reenactment							Cross-id Reenactment	
	L1↓	**PSNR↑**	**SSIM↑**	**FID↓**	**LMD↓**	**AED↓**	**Sync↑**	**FID↓**	**Sync↑**
FOMM[29]	0.046	28.890	0.740	11.04	1.294	0.142	5.17	11.93	3.17
Face-vid2vid [39]	0.062	29.160	0.690	11.47	1.620	0.153	4.96	10.81	4.19
MRAA [30]	0.040	23.351	0.64	11.36	1.280	0.135	3.10	15.61	3.96
PC-AVS [44]	0.081	23.750	0.620	14.32	1.843	0.180	**6.76**	16.78	**6.39**
DA-GAN [12]	0.036	31.220	0.804	9.10	**1.278**	0.129	5.01	9.40	4.71
AVFR-GAN (Ours)	**0.034**	**32.20**	**0.824**	**8.48**	1.280	**0.127**	5.45	**9.05**	4.99

Table 1: Comparison with state-of-the-art methods on Same-identity Reenactment and Cross-identity reenactment on VoxCeleb[22] dataset. ↑ indicates larger is better, and ↓ indicates smaller is better.

4. Experiments and Results

We provide a comprehensive set of evaluations to measure the performance of our proposed method. We perform the quantitative assessment by following the standard benchmarks set by the previous works. We also perform extensive human evaluations to provide a qualitative assessment of the generated results.

Evaluation Set We use the public test set of the VoxCeleb [22] dataset. The dataset contains videos of celebrities. All the videos are preprocessed to 256×256. The test set contains 465 number of videos of different identities making up a total of 76 minutes.

Evaluation Metrics To provide an extensive evaluation of video reconstruction, we use several metrics to measure the performance of different works. We use the following metrics to measure various aspects of our generation. **L1**: It checks the average L1 distance between the generated and ground-truth video. **LMD**: Landmark Distance calculates the distance between detected key points of ground-truth and developed video using a pre-trained facial landmark detector[8]. Please note that this metric was denoted by Average Keypoint Distance in [29]. However, we renamed it Landmark Distance to avoid confusion with the keypoint detector module used in this work. **AED**: Average Euclidean Distance is used to evaluate the identity information. We use Openface[6] to find the feature vectors of generated and ground-truth video and then take the $L2$ distance between them. **PSNR**: Peak Signal to Noise Ratio is used to evaluate the reconstruction quality of the generated image compared to the ground truth image. **SSIM**: Structural Similarity Index evaluates the perceived changes in structural information of an image. We use it along with PSNR as it can also handle global illumination changes. **FID**: Fréchet Inception Distance is used to compare the distribution of generated images with the ground truth image using the features extracted from an InceptionV3 model [32]. **Sync**: Syncnet confidence score is used to measure the amount of lip sync [10].

Comparison with State-of-the-Art Methods We compare our work with the current methods published for the same task. To have a fair comparison, we use the official pre-trained models of FOMM [2], MRAA [3], PC-AVS [5] and DA-GAN [1] from their respective open-source implementations. For Face-Vid2Vid, we use an unofficial implementation in [4]. All the pre-trained models and AVFR-GAN were trained on the same train split and evaluated on the test split of VoxCeleb[22] using two inference strategies defined below.

Same-identity Reenactment We perform the face reenactment task where the source frame and the driving video are of the same person. In this setting, we take the first frame of any video as the source frame and consider the rest of the video as the driving video. The audio chunks corresponding to each driving frame are also fed to the network as input. In this case, we expect the generated output to be as close to the original video as possible. We can therefore calculate metrics like L1, LMD, PSNR, and SSIM, which requires ground truth. We also calculate AED, FID, and Sync metrics for the generated outputs from all the models. From Table 1, it is evident that our method outperforms all the other competing methods. The superior L1 and AED show that our model preserves identity information better. The improvement achieved by our model in terms of LMD indicates the improved structure of generated faces. Interestingly, our model generates improved eye movement in much more detail compared to the previous methods. We got state-of-the-art PSNR, SSIM, and FID scores, correlating with better visual quality. Finally, the sync quality achieved by our algorithm is superior to all the methods except PC-AVS, which performs slightly better in this metric.

Cross-identity Reenactment In this setting, we take a driving video for a different identity and animate a source image. The audio from the driving video is also given as input to the network, as usual. However, since the generated output does not mimic any specific ground truth, we use metrics that do not directly need the same. We use FID, which measures the distance between real and generated

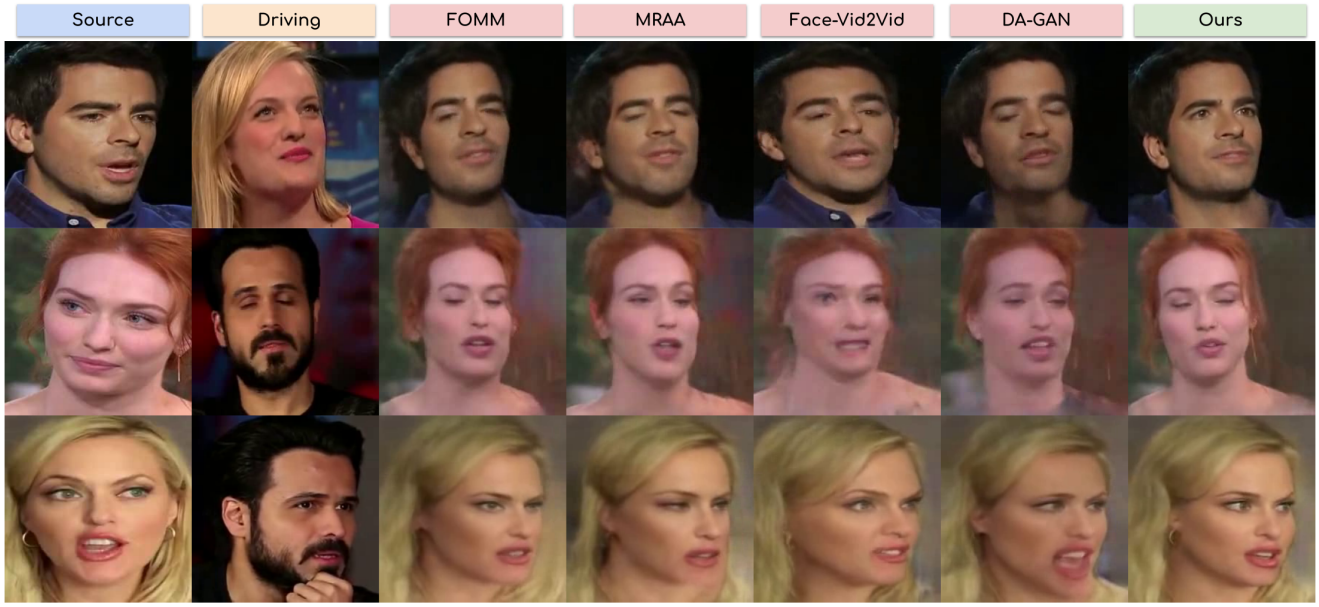

Figure 6: Qualitative comparison on Cross-identity reenactment. Our method gives fewer artifacts, preserves facial structure and handle motion in a better way.

distributions and does not require one-to-one ground truths. We also use Sync to measure the quality of the lip sync in the generated video. As seen in Table 1, we achieve the best FID results and the second-best results in sync trailing only to PC-AVS.

Human Evaluations Since our algorithm generates outputs directly meant for human consumption, we perform extensive human evaluations to ascertain the quality of the generations from our model from a human's perspective. We perform a study enrolling 20 users. Each user is shown generated samples from the state-of-that-art method along with Ours. The users are also shown the source image and the driving video. We select 30 samples from Cross-identity generations. Our user study shows corresponding results from each algorithm side by side, along with the source image and the driving video. The users are asked to rate each generated output based on three characteristics. The users rate the quality of 1. Head pose matching the driving videos, 2. Expressions matching the driving videos, 3. Identity preservation between the source image and the generated videos. The ratings are between 1 to 5, where 1 corresponds to the worst and 5 corresponds to the best. As seen in Table 2, our model consistently yields better results across all the criteria. Our model can enact a better head pose and match expressions of the driving video while preserving the source identities.

	HPMS↑	**EMS↑**	**IPS↑**
FOMM[29]	3.40	3.16	2.80
Face-vid2vid [39]	3.70	3.12	2.66
MRAA [30]	3.26	3.06	2.50
PC-AVS [44]	1.58	1.64	1.92
DA-GAN [12]	3.98	3.82	3.10
AVFR-GAN (Ours)	**4.56**	**4.22**	**3.94**

Table 2: User Study quantitative comparison. 'HPMS' represents Head Pose Matching Score, 'EMS' represents Expression Matching Score and 'IPS' represents Identity Preservation Score. ↑ shows higher is better.

5. Ablation Study

Our proposed approach comprises addition of several key priors and the use of a better image generator. We check the contribution of each of these novel blocks in this section. For setting a baseline (very similar to FOMM), we remove Face Mesh, Face Segmentation, Audio Encoders, and used a basic CNN-based decoder architecture[29, 12, 39]. We add one module at a time to this baseline and train them on the same train-test split. We first add only face mesh and face segmentation to the baseline. We separately also check the effect of adding audio to the baseline. Finally, we combine the structural priors and audio to train a model without the novel identity-aware generator. We calculate SSIM, FID, and Sync metrics and report them in Table 3.

As we observe clearly, the structural priors improve the

Figure 7: Qualitative results on same-identity face reenactment. Upper row: Driving Video, Lower row: Generated Results

	SSIM↑	FID↓	Sync↑
Baseline	0.74	11.04	5.17
+ Structural Prior	0.801	8.98	5.19
+ Audio Prior	0.79	8.69	**5.48**
+ IAG	0.812	8.51	5.13
AVFR-GAN	**0.824**	**8.48**	5.45

Table 3: Ablation Study. The baseline represents the model without face mesh, segmentation, audio, and identity-aware decoder. '+ Structural Prior' represents Baseline with face segmentation and face mesh. '+ Audio Prior' represents Baseline with Audio encoders. '+ IAG' represents Baseline with Identity Aware Generator. ↑ indicates larger is better, and ↓ indicates smaller is better.

SSIM significantly over baseline while audio improves the lip sync quality. We also observe that audio improves the visual quality (measured using FID) of the generations marginally. Finally, the identity-aware face generator gives a significant boost in terms of visual quality improvement.

6. Applications

Our work opens up several applications in the digital industry. Our method can revolutionize multiple industries. We can potentially replace recording famous celebrities in a studio environment costing thousands of dollars; we can animate a single picture of them based on home-recorded driving videos. Similar advances can also be made in the education sector, where online lectures are integral part of education. News readers can reduce their commute and present news from the comfort of their homes by animating their characters. We can also make video calls simpler in more than one way. We can replace the live video feed with a generated one reducing zoom fatigue. More importantly, this can lead to huge bandwidth reduction due to the compact keypoint-based representation, as already noted in [39].

Low-bandwidth Video Conferencing Face reenactment methods can be easily extended for video compression. In the case of a video call between a sender and a receiver, we can first send a single high-resolution frame between the two and follow it up with sending keypoints detected by the keypoint detector for each frame. Our model can then generate the output frames at the receiver's end by considering the high-resolution frame as the source and keypoints from each of the driving frames, similar to the results shown in Figure 7. The 10 keypoints each consist of x and y coordinates and four jacobians, all of which are represented as float values. Therefore, the total bits required to represent a 256×256 frame using FP16 representation is $10 \times 6 \times 16 = 960$ bits. Therefore, the Bits-per-Pixel(BPP) achieved by our model is $\frac{960}{256 \times 256} = 0.014$. We use the latest H.266 codec [7] released in September of 2021 and compress the VoxCeleb test set at the same BPP. While the results generated by our algorithm achieve a FID of 8.48, the H.266 lags by a large margin at 58.32. This indicates the superior quality of the results generated using AVFR-GAN and provides a proof-of-concept for compressing video calls in future work.

7. Further Discussions

In this work, we propose a novel face reenactment network, Audio-Visual Face Reenactment GAN. Our network uses audio-visual cues to reenact a source image according to a driving video. We provide the network with additional structural priors and speech to improve lip synchronization. The final output quality also benefits from a novel identity-aware generator. The improvement in the quality of the generative networks has also led to concerns over its potential misuse. We, therefore, urge the users of any such works to use it ethically. We also encourage users to clearly mark the generated videos with a watermark. We believe these works will benefit and reduce manual effort in professional content creation.

References

[1] Depth-aware generative adversarial network for talking head video generation. https://github.com/harlanhong/CVPR2022-DaGAN1.

[2] First order motion model for image animation. https://github.com/AliaksandrSiarohin/first-order-model.

[3] Motion representations for articulated animation. https://github.com/snap-research/articulated-animation.

[4] One-shot free-view neural talking head synthesis. https://github.com/zhanglonghao1992/One-Shot_Free-View_Neural_Talking_Head_Synthesis.

[5] Pose-controllable talking face generation by implicitly modularized audio-visual representation. https://github.com/Hangz-nju-cuhk/Talking-Face_PC-AVS.

[6] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6(2):20, 2016.

[7] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J. Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (vvc) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021.

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

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

[10] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian conference on computer vision*, pages 87–103. Springer, 2016.

[11] 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.

[12] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3397–3406, 2022.

[13] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14080–14089, 2021.

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

[15] Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1428–1436, 2019.

[16] Lincheng Li, Suzhen Wang, Zhimeng Zhang, Yu Ding, Yixing Zheng, Xin Yu, and Changjie Fan. Write-a-speaker: Text-based emotional and rhythmic talking-head generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1911–1920, 2021.

[17] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Bjorn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10955–10964, 2019.

[18] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (TOG)*, 40(6):1–17, 2021.

[19] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[20] Andrew L. Maas. Rectifier nonlinearities improve neural network acoustic models. 2013.

[21] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[22] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

[23] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.

[24] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV 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*, pages 484–492, 2020.

[25] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.

[26] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016.

[27] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13759–13768, 2021.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[29] 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.

[30] Aliaksandr Siarohin, Oliver Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021.

[31] Yang Song, Jingwen Zhu, Dawei Li, Xiaolong Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786*, 2018.

[32] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[33] Pulkit Tandon, Shubham Chandak, Pat Pataranutaporn, Yimeng Liu, Anesu M Mapuranga, Pattie Maes, Tsachy Weissman, and Misha Sra. Txt2vid: Ultra-low bitrate compression of talking-head videos via text. *arXiv preprint arXiv:2106.14014*, 2021.

[34] Soumya Tripathy, Juho Kannala, and Esa Rahtu. Icface: Interpretable and controllable face reenactment using gans. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[35] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *ArXiv*, abs/1607.08022, 2016.

[36] Lijuan Wang, Wei Han, Frank K Soong, and Qiang Huo. Text driven 3d photo-realistic talking head. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[37] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293*, 2021.

[38] Tianren Wang, Teng Zhang, and Brian Lovell. Faces a la carte: Text-to-face generation via attribute disentanglement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3380–3388, January 2021.

[39] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 10039–10049, 2021.

[40] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9459–9468, 2019.

[41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[42] 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.

[43] 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.

[44] 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.

[45] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020.