Computers & Graphics (2023)

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Visual dubbing pipeline with localized lip-sync and two-pass identity transfer

Dhyey Patel[a], Houssem Zouaghi[b], Sudhir Mudur[a], Eric Paquette[b], Serge Laforest[c], Martin Rouillard[d], Tiberiu Popa[a]

[a]*Concordia University, Montreal*
[b]*École de technologie supérieure, Montreal*
[c]*Audio Z, Montreal*
[d]*Webcargo, Montreal*

ARTICLE INFO

Article history:
Received January 7, 2023

Keywords: Visual dubbing, Reenactment, Style transfer

ABSTRACT

Visual dubbing uses visual computing and deep learning to alter the lip and mouth articulations of the actor to sync with the dubbed speech. It has the potential to greatly improve the content generated from the dubbing industry. Quality of the dubbed result is primary for the industry. An important requirement is that visual lip sync changes be localized to the mouth region and not affect the rest of the actor's face or the rest of the video frame. Current methods can create realistic looking fake faces with expressions. However, many fail to localize lip sync and have quality problems such as identity loss, low-res, blurs, face skin feature or colour loss, and temporal jitter. These problems mainly arise because end-to-end training of networks to correctly disentangle these different visual dubbing parameters (pose, skin colour, identity, lip movements, etc.) is very difficult to achieve. Our main contribution is a new visual dubbing pipeline, in which, instead of end-to-end training we apply incrementally different disentangling techniques for each parameter. Our pipeline is composed of three main steps: pose alignment, identity transfer and video reassembly. Expert models in each step are fine-tuned for the actor. We propose an identity transfer network with an added style block, which with pre-training is able to decouple face components, specifically identity and expression, and also works with short video clips like TV ads. Our pipeline also includes novel stages related to temporal smoothing of the reenacted face, actor specific super resolution to retain fine facial details, and a second pass through the identity transfer network for preserving actor identity. Localization of lip-sync is achieved by restricting changes in the original video frame to just the actor's mouth region. The results are convincing, and a user survey also confirms their quality. Relevant quantitative metrics are included.

1. Introduction

Dubbing is the process of adding or altering speech or other sounds in the audio track of a project that has already been filmed. A major use of dubbing is in movies, serials, advertisements, games, etc., wherein, with the goal of increasing global viewership, the original dialogue is translated into the audience's language of choice, keeping the original actor. Every year, hundreds of films are dubbed into dozens of interna-

e-mail: pateldhyey98@gmail.com (Dhyey Patel), houssem.zouaghi@ymail.com (Houssem Zouaghi), sudhir.mudur@concordia.ca (Sudhir Mudur), eric.paquette@etsmtl.ca (Eric Paquette), serge@audioz.com (Serge Laforest), martin@webcargo.net (Martin Rouillard), tiberiu.popa@concordia.ca (Tiberiu Popa)

tional languages in Hollywood alone. New advancements in visual computing and deep learning such as voice cloning [1] and visual dubbing [2] have the potential to greatly improve the content generated from the dubbing industry [3]. In visual dubbing, lip and mouth configuration of the actor in the original video are also altered to sync with the translated speech, improving viewer experience and speech comprehension [4]. Visual dubbing also makes the translator's task easier when compared with voice only dubbing, a prospect highly welcome to the dubbing industry. This is because the translator has lesser constraints, since lip and mouth articulations of the actor in the original video can be changed to match the translated speech.

Visual dubbing research is mainly spurred by recent advancements combining face tracking and generative networks which have led to the area of "AI deepfakes" and "neural talking heads", with promise in creating realistic fake photos and videos. A common strategy in current work is to use a deep learning network which accepts as input the original video and new dubber audio or video. The network is trained to learn the generation of faces with desired facial expressions including lip and mouth articulations and then used in reenactment of the original actor to mouth the dubbed speech. This reenacted face is then patched back into the original video frame. In principle, these techniques would make it possible to change the actor's lip and mouth configuration as needed. However, many of these take a generic approach, and propose end-to-end networks to learn and transfer face and mouth movements. These solutions often fail to cleanly disentangle the face parameters resulting in error accumulation due to which many a time, one can see that the rest of the face or frame also gets altered, a side effect not acceptable to the industry. Other noticeable quality problems in their output include usually low-resolution, fine facial feature loss, colour inconsistencies, blurring, temporal jitter and actor identity loss.

Our main contribution is a new visual dubbing pipeline, in which, instead of end-to-end training we apply incrementally different disentangling techniques for each parameter. Expert models in each stage are fine-tuned with actor data to maintain identity, mouth expression, fine facial features and resolution.We first identified where the quality problems of color, fine feature loss, temporal jitter, and identity loss get introduced in the output by carefully analyzing each stage of the pipeline. Accordingly, we add appropriate correction actions after each stage. Another point to note is that end-to-end training of large deep neural networks usually require very large datasets, both of source and target actors' speech and facial expressions. This is a problem for the industry, particularly in TV ads which are usually of small duration, like 30 seconds or so, and visual dubbing of supporting cast having only a few lines in a movie. Since we can separately train the different stages in our pipeline we are able to generate good quality visual dubbing outputs even with small training data.

Our visual dubbing pipeline is subject to some hard constrains stemming from industry requirements, specifically that visual lip sync changes should be localised to the mouth and should not affect the rest of the actor's face, like eyes and eyebrows, and certainly not the rest of the video. Another hard requirement is that the original quality of the video in terms of resolution, lighting, colour, background, etc. be retained. We designed our method to comply with these constraints even though in some specific cases this may result in some visual artifacts as shown in the limitations section.

Our innovation is in careful engineering of the dubbing pipeline steps. The input to our pipeline is the actor video and the dubber video. Our strategy is to limit changes in the original frames to just the actor's mouth region, keeping as is, the rest of the actor's face, and rest of the video frame. For this, we generate actor lip and mouth movements mimicking the dubber's speech with the help of our own stylized identity transfer network. Pre-training this network with the CelebA dataset enables it to decouple face components, specifically identity and expression. Additionally, due to this pre-training only a small amount of subject specific training data is needed, typically a 4-5 seconds video is sufficient. Then we replace the mouth region in the original frame with the generated lip and mouth movements. This frame by frame mouth pasting, can sometimes produce mouth quiver and mouth style mismatch. To correct this, we use a second identity transfer pass but this time replacing the dubber video with the reenacted actor video. This helps address the major requirements of identity preservation, lip-sync localization, visual quality retention of the actor's environment, and ability to work with small video clips. Our visual dubbing system ensures high quality through the following: (i) dubber identity does not leak into the generated actor's mouth and face, (ii) mouth patched within the actors' face appears seamless, (iii) temporal stability and (iv) generated actor's mouth has the same quality in resolution colour and lighting. Above (i) and (ii) are achieved by our stylized identity transfer network, and the second identity transfer pass. For (iii) we include two spatio-temporal smoothing steps in different stages to remove temporal jitter. For (iv), since many of our expert models work on lower resolutions, we fine tune a pre-trained super-resolution network to retain intricate face details of the actor such as skin pores, skin colour, lip colour, etc. A user survey confirms the quality of our results. We also present quantitative metrics related to lip sync and overall visual quality.

The rest of the paper is organized as follows. In the next section, we present relevant related work in facial reenactment, neural talking heads and visual dubbing, and contrast these with our system. This is followed by a detailed description of our pipeline including the engineering of various steps that help us address the problems listed above. The need for these steps is substantiated through relevant ablation studies. Then we present various results including comparisons with results from earlier work, when ever possible. The final concluding sections also present the limitations of our system. An accompanying video illustrates our method, ablation study and comparison results. Supplementary material includes example results and our user study questionnaire.

2. Related Work

2.1. Facial Reenactment

Facial reenactment is a conditional face synthesis task which aims to change a target facial expression and pose based on

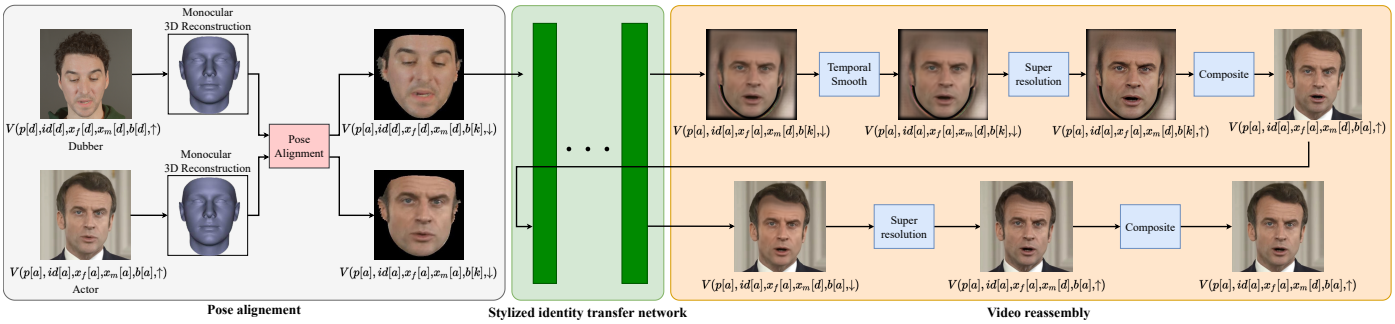

Fig. 1. Our method processes the actor and dubber videos through three main steps. At the core, our stylized identity transfer network changes the face of the dubber to the identity of the actor while preserving the mouth expression. Note that for compactness of visualization we only show the face region, but the method uses the full image frame at every step.

a driving source. Methods can be classified as GAN based or model based. GAN-based methods use image-to-image translation networks [5, 6] for transferring the expression from source to target. They require an intermediate product such as landmarks, dense motion fields or blendshapes. Nirkin et al. [7] use landmarks from the target and train a recurrent neural network for agnostic face swapping and reenactment. Wiles et al. [8] and Siarohin et al. [9] train a network to learn dense motion fields from the source. The target frame is later warped using the learned motion field for pose and expression.Wang et al. [10] propose a few shot video to video synthesis for generating videos using an input semantic map. Facial reenactment using such a method can be achieved by providing an edge map of the source to drive the target generation. Model-based methods use 3D morphable face models (3DMM) [11] to estimate 3DMM parameters. Thies et al. [12] and Ma and Deng [13] conducted an effective deformation transfer to both source and target videos, tracked facial expressions, then re-rendered the synthesized target faces to better fit the retrieved and warped mouth.Kim et al. [14] proposed a method for controlled full head reenactment. They first transfer the head pose, facial expression, and eye motion using 3DMM parameters from source to target. Then train a rendering-to-video translation network to generate photo-realistic output. However, their approach requires large training data, usually hundreds of seconds of video clips. The literature on face reenactment focuses on complete facial expression transfer to the target and work off a globally average expression face for the source. As a result, the mouth expression often lacks the required intensity which is critical in visual dubbing. Moreover, they require a large amount of diverse video footage to train.

2.2. *Talking Heads*

With the recent advancement in deep learning, the problem of talking head generation has enjoyed great success. Talking head generators synthesize an audio-synchronized video given a few facial images for identity using some driving modality, like audio, text or 3DMM parameters. Talking head generation can be subject dependent or independent. Subject dependent methods [15, 16, 17, 18, 19, 20] learn an identity specific embedding using a large subject specific dataset. Using this learned embedding they create photorealistic talking heads of the subject.Nagano et al. [15] and [19] proposed real-time talking head generation techniques for mobile devices. Nagano et al. [15] desgined a network to use facial action units for talking heads generation, whereas [19] suggested to use two stage layered network and reference landmarks. Subject independent methods, also known as few shot generation methods, can work on any identity. Zhou et al. [21] shows pose controllable talking heads for any identity. Han et al. [22] can use either text or audio modalities for realistic talking head generation.Wang et al. [23] and Zakharov et al. [24] use facial keypoints to predict flow to drive a source image using a driving video. Despite these recent breakthroughs, talking head methods cannot be used for dubbing real-world visual content as subject dependent methods would require many hours of video footage for target actor which would be impractical in most cases. On the other hand, subject independent methods result in outputs with unnatural head movements and often lose the identity of the actor as they generate over-smoothed faces.

2.3. *Visual Dubbing*

Based on the modalities used from the source, visual dubbing can be: audio or expression based. Audio-based techniques correct the lip motion of the target to match the source audio. Often these are referred as lip synchronization techniques. Suwajanakorn et al. [25] trained a recurrent neural network to predict mouth shape from raw audio. Based on the generated mouth shape, a realistic texture of mouth was created and composited on target frame. However, they require very large amount of training video, for example, they needed 17 hours of Obama speech footage, making this method impractical in most cases. Audio-based methods [26, 27, 28] can generalize for any identity and voice. Chung et al. [26] jointly trained for audio and video correlation, and they were able to efficiently sync static image with audio but their approach fails for video sequences. Prajwal et al. [27] were the first to propose the use of a powerful lip-sync discriminator with which they achieved good accuracy in syncing an arbitrary video with audio. However, both approaches suffer from blurring of mouth and inconsistent reconstruction of teeth. In comparison, our method produces realistic teeth and has no blurring.

Unlike the above methods which directly morph lips based on audio, Xie et al. [28] proposed to use a two-stage framework.

In the first stage, they train a generator to predict reference face landmarks based on audio. In the second stage, reference landmarks along with the target frame are used to generate the final output. A problem with their approach is that the generated reference landmarks are not always accurate. Their approach also leaks identity. The fundamental problem with audio-based methods is that speech acoustics cannot represent the full range of realistic facial expressions [29] due to which they cannot recreate all possible expressions, say, those present in the actor video input.

Expression-based techniques drive lips in a target video (actor) using expressions from a source video (dubber). Garrido et al. [30] capture facial performance of source and target. They transfer blendshape weights of the mouth from source to target directly. Later, they detect bilabial consonants from audio track and then manually enforce opening and closing of mouth. Finally they render the mouth and composite it on the target frame. Their approach fails to preserve features of the actor while transferring the expression from the dubber. Also, their synthesized inner mouth has observable artifacts and does not look realistic. Kim et al. [31] proposed a style preserving visual dubbing approach. They train a style translation network to learn the mapping of 3DMM expression parameters from source to target. They use a neural face renderer to synthesize a realistic video portrait based on synthesized expression parameters. Their result preserves the style of target mouth. However, their approach ends up manipulating other parts of the face especially the eyes. Successful isolation of expression parameters for monocular face reconstruction is still an unsolved problem which is being extensively investigated. In our visual dubbing pipeline, We do the expression transfer in 2D image space. Suwajanakorn et al. [25] have shown that if source and target are from the same identity, then seemingly realistic expression can be created. Inspired by this, we employ a two pass face-swapping strategy. Our method, while preserving the target identity, accurately transfers lip and mouth movements from dubber to actor. Unlike other known methods, our approach also requires much less data and can work on input videos of only a few seconds.

3. Method

Our goal is to create a video where an actor really looks like speaking in a different language. We take as input two videos A and D: A is the video of an actor uttering some speech in a language L_A and D is a video of a dubber that utters the same speech in a different language L_D. The output is a video R of the actor that appears to be uttering the speech in language L_D, while at the same time maintaining all other visual aspects from the original video: not only the background must remain the same, but it is important that parts of the face, other than lip and mouth, for example the eye region must be kept intact.

To formally express this process we parameterize the videos $V(\cdot)$ by the following parameters: (1) pose p (the rigid transformation of the entire head), (2) identity id (actor or dubber), (3) facial expression x_f (of the actor or of the dubber, excluding mouth), (4) mouth configuration/expression x_m (of the actor or of the dubber), (5) background b (actor video, dubber video or black background) and (6) resolution (high or low). For a compact notation, the values of these parameters are a for actor, d for dubber, $\uparrow$ for high resolution, $\downarrow$ for low resolution and k for videos with a black background.

We introduce the resolution as part of the parameter list because practical dubbing methods should be able to operate on videos of high resolution (1080p or more). Currently, many of the visual dubbing methods that are proposed operate at much lower resolution. We also separate the facial expression from the mouth expression because in a practical commercial context, it is important to maximize the screen real-estate of the original footage, i.e., all other than the mouth region should be kept unchanged from the original video. Therefore:

$$\text{we have } A_0 = V(p[a], id[a], x_f[a], x_m[a], b[a], \uparrow),$$
$$D_0 = V(p[d], id[d], x_f[d], x_m[d], b[d], \uparrow)$$
$$\text{and we want } R = V(p[a], id[a], x_f[a], x_m[d], b[a], \uparrow)$$

The biggest challenge in all visual dubbing methods is how to disentangle and synthesize these parameters. In our method we do not aim at separating them in one end to end neural network; rather we apply incrementally different disentangling techniques for each parameter.

Figure 1 shows the three main steps of our method: (1) *pose alignment* where we register the actor and dubber poses, (2) *identity transfer* where we synthesize the actor with the expression of the dubber and (3) a *video reassembly* step that improves and assembles the final result.

3.1. Pose Alignment

We start by reconciling the pose of the two input videos by rendering them both in the same pose, the pose of the actor, using monocular 3D reconstruction of faces. Various methods have been proposed for 3D facial reconstruction using either parametric models or regression based face trackers. Methods based on parametric reconstruction [32, 33, 34] provide high reconstruction accuracy (NoW Challenge [35]), however they don't reconstruct the inner region of the mouth, which is critical for our application. We would need an extra step for predicting and reconstructing the proxy teeth, and blend them back with reconstructed faces. However, unless the reconstructed proxy is accurate it will lead to uncanny artifacts [30]. Therefore we use the PRNet regression-based face reconstruction method [36]. Using PRNet we obtain $V(p[a], id[a], x_f[a], x_m[a], b[k], \downarrow)$ and $V(p[a], id[d], x_f[d], x_m[d], b[k], \downarrow)$. PRNet provides a rigid 3D transformation to a canonical front facing pose. We use these transformations in order to transform the reconstructed face of the dubber in the same pose as the actor. Even though these videos are rendered at the same resolution as the input video, because of the resampling introduced by the 3D reconstruction and rendering, we technically consider them as low resolution.

The 3D reconstruction being carried out frame by frame could introduce temporal jitter. So, we apply a correction step of temporal smoothing (moving average of 5 frames) on the position of the mesh vertices. As the illumination conditions in the actor and dubber videos are likely to be vastly different, we apply a tonal correction step by shifting the color space [37] of

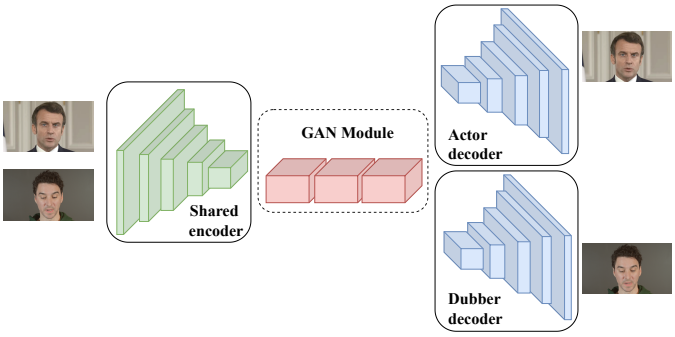

Fig. 2. Our stylized identity transfer network is composed of a shared (actor and dubber) encoder, adapted GAN module, and two decoders (specialized for the actor or dubber).

the dubber video to match that of the actor video. Doing so, also transforms the skin tone of the dubber to that of the actor, which is beneficial in later stages.

3.2. Identity Transfer

Once the videos are aligned, we change the identity of the dubber into the identity of the actor using our stylized identity transfer network to obtain $V(p[a], id[a], x_f[d], x_m[d], b[k], \downarrow)$. We designed a Y-shaped network (Figure 2): single encoder and dual decoder similar to the method of Naruniec et al. [38]. Our goal is to create a network which accurately learns an embedding from a relatively small dataset to synthesize identity transferred faces without changing expression of the target identity. For this we need to decouple facial attribute parameters: identity, pose, and expression. Since, we explicitly rigid align the pose of the actor and dubber, our model only needs to learn to decouple identity and expression in latent space. Recently, StyleGAN [39] has shown unparalleled decoupling of facial parameters. We create our encoder-decoder based on the Xception network [40] and include a GAN module inspired from StyleGan for disentanglement of parameters. Figure 3 shows the full details of this architecture.

Our encoder uses depthwise separable convolution layers like the Xception network. Depthwise separable convolution requires less computational operations and also they provide a dedicated feature pathway for features of high importance. The encoder takes an input image of size $256 \times 256 \times 3$ and calculates a feature map starting from 32 to 1024. The encoder provides two embeddings: one for expression(ϵ_S) and one for identity(ϵ_I), each of size $8 \times 8 \times 1024$. Each decoder takes input map of size $8 \times 8 \times 1024$ and matches corresponding feature level of the encoder.

GAN module first takes expression embedding ϵ_S and passes it through a mapping network of 3 convolutions blocks, each block consisting of convolution, batch norm and leaky relu module. The generated embedding along with identity embedding ϵ_I is passed further through three consecutive blocks, each block having two AdaIN layers. The formulation of AdaIN task can be written as:

$$AdaIN(\epsilon_I, \epsilon_S) = \sigma(\epsilon_S)\frac{\epsilon_I - \mu(\epsilon_I)}{\sigma(\epsilon_I)} + \mu(\epsilon_S) \quad (1)$$

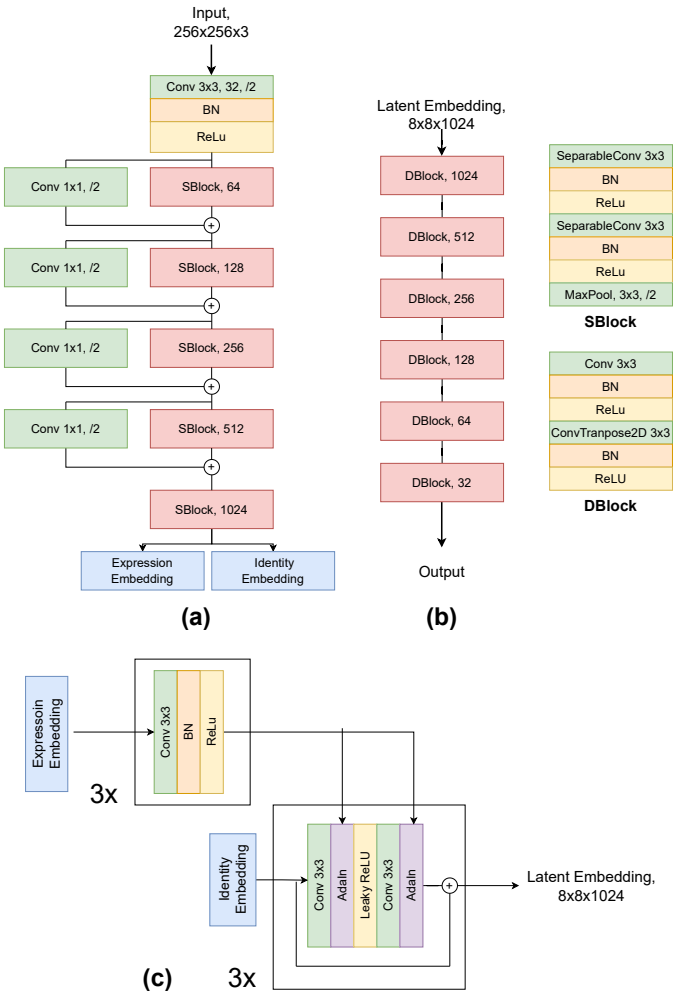

Fig. 3. Detailed architecture of identity transfer network. (a) share encoder (b) decoder (c) gan module

Here μ and σ are the channel wise mean and standard deviation operation.We jointly train the shared encoder and the two decoders during each iteration.

We pre-train our network on the large CelebAHQ dataset [41] for face reconstruction. Pre-training on CelebA helps our encoder and AdaIN blocks to effectively decouple embedding for diverse identity and expression. It also helps our model to converge faster. For subject dependent training we only load weights of encoder and AdaIN block. Decoders start their learning from scratch. This strategy of training a shared encoder with multiple identities for face swapping has been shown to be effective in generating diverse expressions [38]. To obtain the best results, we train our network on the original videos A_0 and D_0, and also on the rendered videos of the actor $V(p[a], id[a], x_f[a], x_m[a], b[k], \downarrow)$ and aligned dubber $V(p[a], id[d], x_f[d], x_m[d], b[k], \downarrow)$. Even though the rendered videos are of a lower quality than the original videos, the black background helps the training to converge faster and with fewer input frames (the network focuses more on the face and less on the background). Faces in the frames are detected cropped and aligned just like typical face recognition pipeline for the training. For losses,we employ *SSIM(Structural Similarity In-*

dex Metric) [42] as facial reconstruction loss. We multiply the input I_{in} and output image I_{out} with the face segmentation mask $\mathbf{m}_f$. Our facial reconstruction loss therefore looks like:

$$\mathcal{L}_{recon} = SSIM(\mathbf{m}_f \bigodot I_{out}, \mathbf{m}_f \bigodot I_{in})$$

Here $\bigodot$ represents element wise multiplication. We also penalize the mouth generation using $\mathcal{L}_2$ pixel-wise reconstruction loss. We use mouth mask $\mathbf{m}_m$ for the mouth segmentation.

$$\mathcal{L}_{mouth} = \|\mathbf{m}_m \bigodot I_{out}, \mathbf{m}_m \bigodot I_{in}\|_2$$

Thus the loss for each common-encoder and decoder pair would be:

$$\mathcal{L}_{actor} = \lambda_1 \cdot \mathcal{L}_{recon} + \lambda_2 \cdot \mathcal{L}_{mouth}$$

$$\mathcal{L}_{dubber} = \lambda_1 \cdot \mathcal{L}_{recon} + \lambda_2 \cdot \mathcal{L}_{mouth}$$

Therefore, our total training loss in the network is:

$$\mathcal{L}_{total} = \mathcal{L}_{actor} + \mathcal{L}_{dubber}$$

For the training we use $\lambda_1 = 4$, $\lambda_2 = 2$. We train our network for 60K iterations, with each original and synthesized dataset, with learning rate $lr = 0.0001$, and it takes about 3 days on a single Nvidia v100 GPU to train. Note that, since our machine has multi-threading issues at that time, causing dataloading bottlenecks, training time may significantly differ if multi-threading was properly supported. During the inference time, we load the weights of shared encoder and actor decoder. We feed the pose aligned dubber video $V(p[a],id[d],x_f[d],x_m[d],b[k],\downarrow)$ to the network to synthesize the reenacted video $V(p[a],id[a],x_f[d],x_m[d],b[k],\downarrow)$ because of which the output in first pass has a black background.

3.3. Video Reassembly

The last step is to superimpose the synthesized part of the face with the original video. As described in Sec. 3.1 the dubber video is rendered in the actor pose using PRNet. PRNet provides the necessary rigid 3D transformations needed for this operation, but unfortunately, the pose estimation is noisy and thus the resulting video can be jittery. Therefore we apply spatio-temporal landmark smoothing to stabilize the video (Section 3.3.1). We further apply a super-resolution step to bring the video to the same resolution as the input (Section 3.3.2) resulting in $V(p[a],id[a],x_f[d],x_m[d],b[k],\uparrow)$. What is left is to cut a mask around the mouth region and composite [43] it with the original actor footage thus obtaining $V(p[a],id[a],x_f[a],x_m[d],b[a],\uparrow)$.

Unfortunately, a simple compositing operation of the mouth region can result in uncanny effects because of positional errors as can be seen in the accompanying video.The positional error occurs when the position of the synthesized mouth doesn't match with the exact position of mouth in the actor. This is because, the 3d rigid transformation from PRNet is sometimes noisy as a result of which the mouth position of pose aligned dubber and actor will differ. The same noisy dubber image when used for synthesizing expression results in incorrect mouth position. An elegant solution to this problem is to use a second pass and recycle the result through the identity transfer step once more with the original A_0 and with $V(p[a],id[a],x_f[a],x_m[d],b[a],\uparrow)$ replacing the dubbing video. This is possible because both these videos are on the same pose.

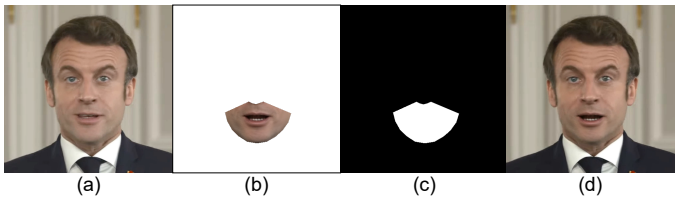

Fig. 4. Frame Compositing: (a) actor frame (b) synthesized expression (c) mask used (d) final result

The resulting output is a convincing dubbing sequence of the actor that retains the content of the original footage everywhere except the mouth region that is lip-synced to the new speech.

This second identity transfer pass as described above is a crucial correction step in our process. This is because of the following. During the first identity transfer the goal is to change the dubber video to have the actor's identity retaining the dubber's facial expressions. Given the small amount of data and especially when there is vast difference in facial features of the dubber and actor, some mix of features is unavoidable. The cut and paste of the mouth region can introduce further positional errors resulting in some uncanny effects. This second pass elegantly corrects these problems using the original actor video.

Figure 4 illustrates the compositing of the synthesized mouth. Figure 4(a) shows the original actor frame. Figure 4(b) shows the synthesized facial expression. Figure 4(c) shows the mask used for blending and Figure 4(d) shows the final result.

3.3.1. Spatio-Temporal Landmark Stabilization

The pose estimation provided by the monocular 3D shape estimation is fairly noisy thus resulting in a jittery video sequence. We extract 2D face landmarks using DLib [44] for the reenacted video $V(p[a],id[a],x_f[a],x_m[d],b[k],\downarrow)$, and temporally smooth their respective positions. We then compute a 2D rigid alignment to align the reenacted video landmarks to the smoothed landmarks (applying that transformation to the video). The most difficult challenge in temporal stabilization is the balance between jitter and lag [45]: during the parts of the video where the scene is mostly static the jittering artifact is more prominent and thus a larger smoothing window needs to be applied while during the parts where there is fast movement, too much smoothing might result in video lag. To balance these we employ the 1 Euro filter proposed by Casiez et al. [45].

3.3.2. Finely Tuned Super Resolution

We use the network of Yang et al. [46] to do a super resolution of 512×512 from our synthesized output of 256×256. One challenge with any super-resolution method is that it will always make the result visually sharper and clearer. However, given a video input, due to motion, focusing and other reasons, the input footage might not always be sharp and the super-resolution image might not, in fact, match the quality of the

Fig. 5. Finely Tuned Super Resolution. (a) input frame (b) output before super resolution (c) super resolution using only Yang et al. [46] (d) super resolution using Yang et al. [46] and our data augmentation.

input. We address this using data augmentation. We take the high-resolution video input and create a low-resolution version by blurring it. We blur the video in two different ways: 1) using the guassian blurring from the above paper, 2) we self-swap the original actor video(passing the actor image from shared encoder-actor decoder pair) using our trained identity transfer network. We use these pairs of images of the actor to further train the super-resolution network. We finetune the model for 1000 iterations. It takes about 35 minutes for finetuning on single Nvidia Tesla v100 GPU. During finetuning we change following parameters compared to the ones in the original paper: learning rate $lr = 0.0002$, $\alpha = 1.0$ and $\beta = 1.0$ This fine tuning step yielded overall better results.

We illustrate this in Figure 5. Figure 5(a) shows the input frame. Figure 5(b) shows the output before the super resolution, where it is clear that the visual quality is lower. We purposefully chose a frame where the input frame and the output frame happen to have similar mouth opening. Figure 5(c) shows the frame after using only Yang et al. [46]. Figure 5(d) shows the final output frame using Yang et al. [46] with our data augmentation. The final result matches the quality of the input frame much better: it is not as blurry as the output without super resolution but not as sharp as the super resolution without our data augmentation.

4. Results

We tested our method on over a dozen production quality dubbing scenarios provided by our industry collaborators. Because of copyright reasons, here, we can only show results on copyright free YouTube video clips. For original videos (i.e. actors) we selected a few clips of speeches of international public figures and the dubbing was done by professionals. These dubbers have training in traditional dubbing, but no additional training was done for these dubbing sessions other than mentioning that in addition to the sound recording there is a camera that records their facial expression. In fact, as you can see in the accompanying video, the dubbers actually keep their heads quite down, making it rather challenging for face transfer. In addition to these sequences we also extracted a few video sequences from the supplementary material submitted to the ACM digital library of Kim et al. [31] and compared our results to theirs. All the results generated by us are available as a video in the supplemental material. We put all the necessary videos (actor, dubber and result) at full resolution and length. Comparative results as well as ablation studies are included in the supplemental video.

5. Evaluation

Using our selected clips of public figures, we evaluate our method against the recent state-of-the-art methods - audio based lip-sync [27], talking head generation [47] and face reenactment [9]. For visual dubbing methods, we found it challenging to evaluate our method against other methods. There are no uniform datasets available targeted to this problem, and neither is the source code of these methods. Still from the recent video dubbing methods, we cropped the actor and dubber sequences from the online video presentation of Kim et al. [31] and we processed them with our pipeline. This comparison is not an apple to apple comparison because we do not have the same training videos in terms of both length and resolution. Nevertheless, even with much less training data (185 to 300 frames for our approach, 3000 to 9000 for Kim et al. [31]), our results are comparable and in some respects better than those of Kim et al. [31].

We also compared our results to traditional dubbing through a user study where we ask a few qualitative questions related to the lip synchronization with the audio, visual quality of the face, and blind selection of preferred video out of our results and traditional dubbing.

5.1. Quantitative Evaluation

For quantitative comparisons, we evaluate lip-sync accuracy and visual quality of the generated results. We use the Landmark Distance (LMD) metric proposed by Chen et al. [48]. LMD calculates normalized euclidean distance between landmarks of the mouth for the generated result and the dubber on per frame basis. Our use of LMD is justified because of the availability of ground-truth expression from the paired actor-dubber sequence. Table 1 shows LMD score on our dataset between our methods and various state-of-the-art methods. We also employ Learned Perceptual Image Patch Similarity (LPIPS [49]), Frechet Inception Distance (FID [50]) and Peak Signal to Noise Ratio (PSNR) metrics to assess the overall visual quality of the generated result compared to the original actor video. Table 2 shows the result of the visual quality evaluation.

For the videos from Kim et al. [31], we calculate lip-sync accuracy as well as visual metrics (Table 3). Since the cropped videos were of low resolution, visual metrics comparison in this scenario might not be completely accurate.

Our method largely outperforms all methods in FID and PSNR, and it outperforms FOMM and MakeItTalk in LPIPS. When compared to Wav2Lip using the LPIPS metric as well as when using the LMD metric, the values are similar. However, a major advantage of our method is that it uses much less training data compared to all the above methods and it is specifically designed to keep the original video intact in the upper face region, which if changed can lead to uncanny effects [51].

Table 1. Landmark Distance (LMD) metric (lower is better) of mouth expression between our generated result and the dubber video computed on our dataset.

Dataset	Ours	Wav2Lip	FOMM	MakeItTalk
Macron	**0.78**	0.93	0.81	0.79
Kovind EN	**0.68**	0.72	0.69	0.75
Kovind FR	0.73	0.74	0.79	**0.67**

Table 2. Visual quality evaluation (LIPS - lower is better, FID - lower is better, PSNR - higher is better) of results generated using other methods compared to ours.

		Metrics		
Dataset	Methods	LPIPS ↓	FID ↓	PSNR ↑
Macron	Ours	**0.02**	**1.04**	**37.13**
	Wav2Lip	0.04	4.77	32.30
	FOMM	0.13	11.50	16.01
	MakeItTalk	0.12	16.45	16.73
Kovind En	Ours	**0.04**	**5.30**	26.12
	Wav2Lip	0.05	11.55	**26.15**
	FOMM	0.12	60.60	19.58
	MakeItTalk	0.11	90.99	20.65
Kovind FR	Ours	**0.04**	**5.51**	**26.14**
	Wav2Lip	0.05	11.63	26.09
	FOMM	0.12	52.31	19.48
	MakeItTalk	0.12	76.41	20.77

5.2. User Study

The goal of the visual dubbing process is, ultimately, to produce video content that yields a superior experience to a viewer. Obtaining real-world opinions is very important since human eyes are the expert discriminator for assessing any audio-visual mismatch and temporal artifact. Accordingly, we carried out a user study that compares our visual dubbing method to the current dubbing method used in the industry which is overlaying the audio content. Using three of our result videos, we showed our participants the result of our method and the result of a professionally produced (sound only) dubbing sequence, and asked a few questions regarding the videos. Our study has 23 participants. The questionnaire they answered is provided in the supplementary material and the three videos shown correspond to the results 1-3 in our supplementary material results video. We summarize the findings in Table 4. A vast majority (69.6% to 87.0%) of the participants found the lip motions in our results to be synchronized with the audio. Moreover, when asked about the overall visual quality of the face in our results, 60.8% to 82.6% of the participants said our results are good or excellent. Finally, when presented with clips which were traditionally dubbed and our visual dubbing results, a majority (65.2% to 91.3%) preferred visual dubbing results.

Table 3. Comparison with Kim et al. [31]. LMD - lower is better, LIPS - lower is better, FID - lower is better, PSNR - higher is better.

		Metrics			
Dataset	Methods	LMD ↓	LPIPS ↓	FID ↓	PSNR ↑
Obama	Ours	**0.56**	0.02	5.44	**34.45**
	Kim [2019]	0.67	**0.02**	**2.74**	30.16
E. and J.	Ours	**1.16**	**0.03**	**22.39**	**36.32**
	Kim [2019]	1.24	0.06	24.39	26.90
F. and K.	Ours	**1.16**	**0.17**	**31.50**	16.18
	Kim [2019]	1.52	0.18	36.45	**16.74**

Table 4. User study statistics (in percentage). For "Lip Sync Acc." we asked the participants if the lips motion is synchronized with the audio (Strongly agree, Agree, Neutral, Disagree, Strongly Disagree) and here we report "Strongly agree"+"Agree". For "Visual Quality" we asked the participants to rate the overall visual quality (Excellent, Good, Fair, Poor, Very poor) and here we report "Excellent"+"Good". "Preference" reports which clip the participant preferred. The values do not sum to 100% as there was a "No preference" option.

			Preference	
Dataset	Lip Sync Acc.	Visual Quality	Visual Dubbed	Traditional Dubbed
Macron	78.2%	82.6%	91.3%	4.3%
Kovind EN	69.6%	60.8%	65.2%	17.4%
Kovind FR	87.0%	73.9%	82.6%	8.7%

6. Discussion

For the spatio-temporal stabilization step, even if the mouth alignment is not completely perfect after the stabilization, the second pass through the stylized identity transfer network fine tunes the alignment of the mouth and face very well. We also tested doing a 2D alignment of the reenacted video toward the original video of the actor. That proved to introduce some more jittering as the chin and mouth movements of the original actor's video are not the same as the ones of the reenacted video.

Our second identity transfer pass elegantly solves the uncanny effect resulting from first cut and paste. However, it slightly diminishes the expression dynamics in the final generated output.

Limitations. Our most important limitation stems from the requirement that we composite only the lower part of the face resulting in occasional artifacts. Our strategy to repeat the face swap process improves the results, but not in all cases. More specifically, in frames where the pose of the actor frame has the jaw open while the dubber has it closed, the pasting results in a double chin artifact that can be seen in the video comparison with Kim et al. [31]. There are also special cases when moving wrinkle artifact is observed with actor having strong nasolabial folds. This occurs when the part of the nasolabial fold, constructed in the synthesized mouth, does not completely align with the one in upper part of the face.

Another limitation comes from the automatic landmark identification. Sometimes we had to manually adjust the landmarks of the actor or dubber. For example, the Macron and Merkel dubbing worked fine with the automatic landmarks, but the Kovind dubbing required the adjustment of landmarks for 200 out of 1185 frames of the dubber. This is solely because of the landmark detector of DLib [44]. Our method is independent of that detector and can benefit from more precise landmark detection methods, present or future.

Ethics considerations. Our method, like dozens of other deep fake methods can be misused for malicious purposes such as misrepresenting individual and spreading misinformation. We are mitigating these risks by only providing the code to professional dubbing companies that have a transparent professional conduct.

7. Ablation Studies

In the accompanying video we have included a section with four ablation studies that we performed in order to demonstrate the necessity of individual steps of our method. In the first one, we look at the impact of the second identity transfer pass. As can be seen in the video, without this pass the result exhibits various temporal artifacts introduced when pasting only the mouth region and sometimes further exacerbated by the Poisson blending step. The second ablation study highlights the necessity of the temporal stabilization steps. In the third ablation study, we vary the footage available for training to demonstrate our claim of not requiring large training data. We compare the original result that was trained on 30 seconds to a sub-clip trained only on 5 seconds. As we can see, in general, there are only minor visual changes between them illustrating the fact that our method works well even on short clips. Finally, we show the impact of the pre-training and AdaIN blocks in the stylized identity transfer network. For short clips, especially, where the lack of training data is an issue, this step is very important.

8. Conclusion

In this work we present a new visual dubbing pipeline where the main design objectives, raised from typical industry scenarios, are the preservation of the rest of the face expression from the original actor footage, the ability to deliver good results on short video clips, and maintaining the resolution and general visual quality of the input. The pipeline design evolved over a continual improvement process in which our industry collaborators provided us actor (real TV ads) and dubber videos, and feedback on output from our pipeline that drove the changes in the pipeline steps in an iterative fashion.

Our pipeline contains several novel ideas and techniques such as a two-pass identity transfer, temporal stabilization, data augmentation for both identity transfer as well as fine-tuned super resolution. The pipeline enables us to disentangle the different parameters in visual dubbing using a step-wise approach, something which is difficult to achieve using end-to-end trained networks.

We evaluate our method qualitatively as well as quantitatively on professionally produced dubbing clips showing the real-world potential of our pipeline. Our results are convincing and confirmed by a user study focused on the overall experience of the dubbing results.

Acknowledgments

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). This work was supported by Mitacs through the Mitacs Accelerate program. We would also like to thank AudioZ for facilitating the data capture and providing various support for this project. We would like to thank the anonymous reviewers for their vital feedback.

References

[1] Arık, SO, Chen, J, Peng, K, Ping, W, Zhou, Y. Neural voice cloning with a few samples. In: NIPS. Red Hook, NY, USA: Curran Associates Inc.; 2018, p. 10040–10050.

[2] Yang, Y, Shillingford, B, Assael, YM, Wang, M, Liu, W, Chen, Y, et al. Large-scale multilingual audio visual dubbing. CoRR 2020;abs/2011.03530. URL: `https://arxiv.org/abs/2011.03530`. `arXiv:2011.03530`.

[3] Mukherjee, S. Now the voice dubbing industry is being disrupted by ai. `https://analyticsindiamag.com/now-the-voice-dubbing-industry-is-being-disrupted-by-ai/`; 2022.

[4] Begau, A, Klatt, LI, Wascher, E, Schneider, D, Getzmann, S. Do congruent lip movements facilitate speech processing in a dynamic audiovisual multi-talker scenario? an erp study with older and younger adults. Behavioural Brain Research 2021;412:113436.

[5] Zhu, JY, Park, T, Isola, P, Efros, AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE/CVF ICCV. Los Alamitos, USA: IEEE; 2017, p. 2242–2251.

[6] Isola, P, Zhu, JY, Zhou, T, Efros, AA. Image-to-image translation with conditional adversarial networks. In: CVPR. Los Alamitos, USA: IEEE/CVF; 2017, p. 5967–5976.

[7] Nirkin, Y, Keller, Y, Hassner, T. Fsgan: Subject agnostic face swapping and reenactment. In: IEEE/CVF ICCV. Los Alamitos, USA: IEEE; 2019, p. 7183–7192.

[8] Wiles, O, Koepke, AS, Zisserman, A. X2face: A network for controlling face generation using images, audio, and pose codes. In: ECCV. Cham: Springer; 2018, p. 690–706.

[9] Siarohin, A, Lathuilière, S, Tulyakov, S, Ricci, E, Sebe, N. First order motion model for image animation. In: Wallach, H, Larochelle, H, Beygelzimer, A, d'Alché-Buc, F, Fox, E, Garnett, R, editors. NIPS; vol. 32. Vancouver, BC, Canada: Curran Associates, Inc.; 2019,.

[10] Wang, TC, Liu, MY, Tao, A, Liu, G, Kautz, J, Catanzaro, B. Few-shot video-to-video synthesis. arXiv preprint arXiv:191012713 2019;.

[11] Blanz, V, Vetter, T. A morphable model for the synthesis of 3d faces. In: ACM SIGGRAPH. SIGGRAPH '99; USA: ACM. ISBN 0201485605; 1999, p. 187–194.

[12] Thies, J, Zollhofer, M, Stamminger, M, Theobalt, C, Nießner, M. Face2face: Real-time face capture and reenactment of rgb videos. In: IEEE CVPR. Las Vegas, USA: IEEE; 2016, p. 2387–2395.

[13] Ma, L, Deng, Z. Real-time hierarchical facial performance capture. In: Proc. of ACM SIGGRAPH I3D. New York, USA: ACM. ISBN 9781450363105; 2019,.

[14] Kim, H, Garrido, P, Tewari, A, Xu, W, Thies, J, Niessner, M, et al. Deep video portraits. ACM Transactions on Graphics (TOG) 2018;37(4):1–14.

[15] Nagano, K, Seo, J, Xing, J, Wei, L, Li, Z, Saito, S, et al. pagan: real-time avatars using dynamic textures. ACM Trans Graph 2018;37(6):258–1.

[16] Ji, X, Zhou, H, Wang, K, Wu, W, Loy, CC, Cao, X, et al. Audio-driven emotional video portraits. In: IEEE/CVF CVPR. Los Alamitos, CA, USA: IEEE Computer Society; 2021, p. 14080–14089.

[17] Lu, Y, Chai, J, Cao, X. Live Speech Portraits: Real-time photorealistic talking-head animation. ACM Trans Graph 2021;40(6).

[18] Thies, J, Elgharib, M, Tewari, A, Theobalt, C, Nießner, M. Neural voice puppetry: Audio-driven facial reenactment. In: ECCV. Cham: Springer; 2020, p. 716–731.

[19] Zakharov, E, Ivakhnenko, A, Shysheya, A, Lempitsky, V. Fast bi-layer neural synthesis of one-shot realistic head avatars. In: European Conference on Computer Vision. Springer; 2020, p. 524–540.

[20] Fried, O, Tewari, A, Zollhöfer, M, Finkelstein, A, Shechtman, E, Goldman, DB, et al. Text-based editing of talking-head video. ACM Trans Graph 2019;38(4):68:1–68:14.

[21] Zhou, H, Sun, Y, Wu, W, Loy, CC, Wang, X, Liu, Z. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: CVPR. Nashville, TN, USA: IEEE; 2021, p. 4174–4184.

[22] Han, L, Ren, J, Lee, HY, Barbieri, F, Olszewski, K, Minaee, S, et al. Show me what and tell me how: Video synthesis via multimodal conditioning. arXiv 2022;abs/2203.02573.

[23] Wang, TC, Mallya, A, Liu, MY. One-shot free-view neural talking-head synthesis for video conferencing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, p. 10039–10049.

[24] Zakharov, E, Shysheya, A, Burkov, E, Lempitsky, V. Few-shot adversarial learning of realistic neural talking head models. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 9459–9468.

[25] Suwajanakorn, S, Seitz, SM, Kemelmacher-Shlizerman, I. Synthesizing obama: Learning lip sync from audio. ACM Trans Graph 2017;36(4).

[26] Chung, JS, Jamaludin, A, Zisserman, A. You said that? International Journal of Computer Vision 2017;127:1768–1779.

[27] Prajwal, KR, Mukhopadhyay, R, Namboodiri, VP, Jawahar, C. 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; New York, NY, USA: ACM. ISBN 9781450379885; 2020, p. 484–492.

[28] Xie, T, Liao, L, Bi, C, Tang, B, Yin, X, Yang, J, et al. Towards realistic visual dubbing with heterogeneous sources. In: Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event, China: ACM; 2021, p. 1739–1747.

[29] Yehia, H, Rubin, P, Vatikiotis-Bateson, E. Quantitative association of vocal-tract and facial behavior. Speech Communication 1998;26(1-2):23–43.

[30] Garrido, P, Valgaerts, L, Sarmadi, H, Steiner, I, Varanasi, K, Pérez, P, et al. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. Comput Graph Forum 2015;34(2):193–204.

[31] Kim, H, Elgharib, M, Zollhöfer, M, Seidel, HP, Beeler, T, Richardt, C, et al. Neural style-preserving visual dubbing. ACM Trans Graph 2019;38(6):1–13.

[32] Deng, Y, Yang, J, Xu, S, Chen, D, Jia, Y, Tong, X. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019, p. 0–0.

[33] Guo, J, Zhu, X, Yang, Y, Yang, F, Lei, Z, Li, SZ. Towards fast, accurate and stable 3d dense face alignment. In: European Conference on Computer Vision. Springer; 2020, p. 152–168.

[34] Feng, Y, Feng, H, Black, MJ, Bolkart, T. Learning an animatable detailed 3d face model from in-the-wild images. ACM Transactions on Graphics (ToG) 2021;40(4):1–13.

[35] Sanyal, S, Bolkart, T, Feng, H, Black, M. Learning to regress 3D face shape and expression from an image without 3D supervision. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2019, p. 7763–7772.

[36] Feng, Y, Wu, F, Shao, X, Wang, Y, Zhou, X. Joint 3d face reconstruction and dense alignment with position map regression network. In: Proceedings of the European conference on computer vision (ECCV). Cham: Springer; 2018, p. 534–551.

[37] Reinhard, E, Adhikhmin, M, Gooch, B, Shirley, P. Color transfer between images. IEEE Computer graphics and applications 2001;21(5):34–41.

[38] Naruniec, J, Helminger, L, Schroers, C, Weber, R. High-Resolution Neural Face Swapping for Visual Effects. Computer Graphics Forum 2020;39(4):173–184.

[39] Karras, T, Laine, S, Aila, T. A style-based generator architecture for generative adversarial networks. In: IEEE/CVF CVPR. Los Alamitos, USA: IEEE; 2019, p. 4401–4410.

[40] Chollet, F. Xception: Deep learning with depthwise separable convolutions. In: CVPR. Los Alamitos, CA, USA: IEEE; 2017, p. 1800–1807.

[41] Karras, T, Laine, S, Aila, T. A style-based generator architecture for generative adversarial networks. In: IEEE/CVF CVPR. Los Alamitos, CA, USA: IEEE; 2019, p. 4396–4405.

[42] Wang, Z, Bovik, AC, Sheikh, HR, Simoncelli, EP. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 2004;13(4):600–612.

[43] Pérez, P, Gangnet, M, Blake, A. Poisson image editing. ACM Trans Graph 2003;22(3):313–318.

[44] King, DE. Dlib-ml: A machine learning toolkit. J Mach Learn Res 2009;10:1755–1758.

[45] Casiez, G, Roussel, N, Vogel, D. 1 € filter: A simple speed-based low-pass filter for noisy input in interactive systems. In: CHI'12, the 30th Conference on Human Factors in Computing Systems. CHI '12; New York, NY, USA: ACM. ISBN 9781450310154; 2012, p. 2527–2530.

[46] Yang, T, Ren, P, Xie, X, Zhang, L. Gan prior embedded network for blind face restoration in the wild. In: IEEE/CVF CVPR. Nashville, TN, USA: IEEE; 2021, p. 672–681.

[47] Zhou, Y, Han, X, Shechtman, E, Echevarria, J, Kalogerakis, E, Li, D. Makelttalk: speaker-aware talking-head animation. ACM Transactions on Graphics (TOG) 2020;39(6):1–15.

[48] Chen, L, Li, Z, Maddox, RK, Duan, Z, Xu, C. Lip movements generation at a glance. In: ECCV. Cham: Springer; 2018, p. 520–535.

[49] Zhang, R, Isola, P, Efros, AA, Shechtman, E, Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. Los Alamitos, USA: IEEE/CVF; 2018, p. 586–595.

[50] Heusel, M, Ramsauer, H, Unterthiner, T, Nessler, B, Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems 2017;30:6629–6640.

[51] Kuster, C, Popa, T, Bazin, JC, Gotsman, C, Gross, M. Gaze correction for home video conferencing. ACM Transactions on Graphics (TOG) 2012;31(6):1–6.