FaceOff: A Video-to-Video Face Swapping System (Supplementary)

Aditya Agarwal* IIIT Hyderabad

aditya.ag@research.iiit.ac.in

Rudrabha Mukhopadhyay IIIT Hyderabad

radrabha@research.iiit.ac.in

Bipasha Sen* IIIT Hyderabad bipasha.sen@research.iiit.ac.in

Vinay Namboodiri University of Bath vpn22@bath.ac.uk C V Jawahar IIIT Hyderabad jawahar@iiit.ac.in

1. Network Design

We adopt the architecture of VQVAE2 [2]. VQVAE2 encodes the input into multiple hierarchies: top and bottom. We adopt the same architecture but modify it in two fundamental ways. (1) VQVAE2 is an autoencoding network and thus computes the distance between the input and the output of dimension $H \times W \times C$ – the height, the width of the image, and the number of input channels, respectively. In our case, input is a channel-wise concatenation of the source foreground, f_{s_i} , and target background, b_{t_i} , giving a dimension of $H \times W \times 6$, and thus, the output generated by our network is of the same dimension $H \times W \times 6$. During training, instead of the input, we compute the loss against the ground truth video, s_i , of dimension $H \times W \times 3$. Thus, we only consider the first three channels of $H \times W \times 6$ output at the network's output. Similarly, we only consider the first three channels as our output at the inference. (2) VQVAE2 operates at a frame level and thus cannot model temporal properties. Thus, we add temporal modules in the network just before the quantization block. At each hierarchy, the encoder produces a latent of dimension $(B * T) \times C \times H \times W$. Here, we expand the batch dimension to convert the flattened input into videos. These video latents of dimension $B \times T \times C \times H \times W$ are then passed through the temporal block made of 3D convolution and ReLU layers (see Fig. 2, main paper). Post this step, we again convert the batch dimension to (B * T). The losses are then applied frame-by-frame. The temporal layers learn to identify the properties across the video and produce a blended encoding even with a frame-by-frame loss. At this point, the encoder outputs are quantized, and we adopt the decoder architecture of VQVAE2 for decoding the latent.

1.1. Our results and Potential Applications

Our approach has several potential applications, especially multimedia, entertainment, and education.

We demonstrate two such applications in this paper. The first is depicted in Fig. 6 of the main paper, which shows a real-use case of Paul Walker. In post-production, the VFX team replaced the face of Cody and Caleb Walker, who acted as Paul's double¹. The team underwent extensive graphical post-processing to superimpose Paul's face from previous recordings of Cody and Caleb. In Fig. 1, we demonstrate another result of FaceOff. Here, we simulate a scenario of body doubles. Nolan, the actor in the source video, is 'working from home' recording his dialogues and expressions at the convenience of his home. Joey Tribiani, the double in the target video, acts in the famous sitcom FRIENDS. FaceOff swaps Nolan into the scene in one forward pass! We show such an application in the supplementary video and encourage our readers to view the result of double-actor V2V face-swapping. FaceOff can potentially save millions of dollars and reduce months of postproduction edits to merely a few minutes of touch-ups on top of the FaceOff output!

Another application of our work is post-production movie editing. Today, multiple scenes are anticipated in advance to avoid retakes during post-production. Our work will encourage the movie-production team to become more flexible with doubles and post-production movie edits.

FaceOff also has huge potential in the advertisement sector and could be a potential futuristic technique for making advertising videos. Today, the VFX and CGI take abundant resources for V2V face swapping, whereas, with our work, one could replace themselves in a sitcom in less than a second. This could also become a potential teaching technique. For example, creating light-hearted advisory videos about vital life lessons for students. Our work can also be

^{*}Equal contribution

¹Redevelopment of Walker's character



Figure 1: Limitations of our approach. Artifacts such as hair strands and spectacles are visible. In case of extreme pose change, the network struggles to produce a coherent output.

applied in animation² to swap an existing face/background in multiple scenes.

2. Limitations

Our work fundamentally lacks two areas: (1) Pose difference in the 'Z' direction (normal to the image) between the source and target. The network struggles to generate coherent outputs. As can be seen in Fig. 1, the lips and the overall production seem unnatural. Going beyond 2D images and exploring the space of 3D modeling could be an exciting way to approach this issue. (2) Difference in face ornaments. As can be seen in Fig. 1, artifacts such as part of the hair and spectacles are visible in the output. As we avoid adding a discriminator, the model does not learn to 'remove' any input part to make the output more realistic. For future work, one could experiment with soft discriminators such that there are minimum hallucinations.

Lastly, we extract the source face using the eye and mouth region landmarks. However, a part of one's identity also includes the head region. We do this to preserve the pose of the target. In the specific use-case we tackle, a double is selected such that the head of the double is similar, if not the same, to the actor (see Fig. 1, main paper). Thus, extracting only the face region is sufficient for preserving the identity in our case. However, to preserve the entire identity, one would have to move from face-swapping to head-replacement [1], which would also be an interesting direction of exploration. Here, one would need to be able to transfer the head pose of the target to the source head while preserving the other necessary characteristics.

3. Ethical Issues

Unlike other generative works in similar settings, we do not re-enact a given identity according to a driving video. Our work focuses on swapping relevant parts of the source video onto the target video so that the expression and lip movements of the source video are preserved. At the same time, the head motion and background remain the same as the target video. This ensures that the generated identity and the spoken content in the generated video match the source speaker (extensively evaluated in Table 2, main paper). Thus, body doubles and doppelgangers of celebrities cannot be directly used to re-enact a target celebrity video since the final generated identity will be copied from the source. However, since our work deals with modifying critical facial features of the target identity, we decide to take further steps to ensure fair use. We will only release the code after signing legal agreements with the users to maintain records. We will also use a visible watermark on the generated video to ensure they remain identifiable and fake.

4. Experimental Setup

4.1. Hardware Setup

All of our models are trained and inferred on NVIDIA GTX 3080 Ti using 4 GPUs and 1 GPU respectively.

4.2. Dataset

	Name	Nationality	YouTube Channel
1.	Anfisa Nava	Russia	ANFISAofficial
2.	Sejal Kumar	India	sejalkumar7theclothingedit
3.	Johnny Harris	USA	johnnyharris
4.	BestDressed	USA	bestdressed
5.	Jack Edwards	UK	thejackexperience

Table 1: Speakers in the training dataset collected from publicly available YouTube VLOG videos.

To create the **training** dataset, we curate publicly available unconstrained YouTube VLOG videos. It includes five different YouTubers; the specifications of the same are provided in Table 1. The data amounts to a total of <u>15 hours</u> of video divided equally among all speakers. All the speakers speak in English, although they have different accents based on their nationality. The details of the videos, along with the timestamp, will be released publicly to promote future research.

The **test** set is also curated from unconstrained YouTube videos. The videos have a different identity, background, and light setting from the training set. Furthermore, they are selected from a widely varying timeline ranging from the 1990s to the late 2021s! This ensures we cover different video capture technologies, compression techniques,

²List of recycled animation in Disney movies



Figure 2: Ablation Experiments. In each of the experiment, we remove the type of error mentioned at the time of selfsupervised training. Here, we present the results of the trained models at the inference on cross-identity.

etc. Specifically, the videos are collected from Sitcom snippets, interviews, and movies. Some examples are The Office (Sitcom), Alex Honnold's Interviews, Think Media's tutorials, and FRIENDS (Sitcom).

4.3. Human Evaluation

We conduct human evaluations as part of our qualitative evaluations, primarily to assess the quality of videoto-video face swapping achieved by our network. We randomly select ten videos from our curated dataset, and the results from all the comparisons and our network are displayed in a random order to the user. A pool of 50 participants is asked to assign a score between 1-10, indicating the perceptual quality of the generated videos. Our participant pool comprised people aged between 25-45 years of age. At the time of rating, every user was asked to rate a video on a scale of 1-10, 1 and 10 being the worst and the best, respectively. Each user was shown a source video, a target video, and the final swapped video. The swapped video could be randomly from FSGAN, Motion-coseg, or FaceOff. Each user saw 10 instances of each category during rating. They had to answer the following three questions: (1) How natural does this video (swapped) look? (2) How similar is the expression in the swapped video to the source expression? and (3) How similar is the identity in the swapped video to the source identity? No additional directions were given to the users for rating. Along with the rating, they were also asked to submit their subjective opinion on the naturalness aspect of the swapped video. The mean opinion scores of all the users are reported in the main paper. We also try to summarize their opinion in this section.

As was observed in Table 2 of the main paper, we outperformed the existing approaches in preserving the source identity in both quantitative and qualitative evaluation. However, FSGAN was voted slightly better qualitatively for the naturalness factor. Hereon, we will discuss the naturalness factor of the observed videos. Out of the three, the highest variations in the user rating were observed to be in Motion-coseg. FaceOff had the least variation in rating, and almost all the videos appear natural. Although FS-GAN was rated highest in terms of naturalness, the users commented that the output had unnatural color. Despite the drawback, the users agreed that the overall expression and the swapped person looked natural. It is to be noted that despite FSGAN being voted as producing more natural outputs than FaceOff, the task of identity swapping was unanimously voted to be superior in FaceOff. Although FSGAN preserved the source identity and looked more natural, the users agreed that the output had a little match with either of the expressions - source or target. This meant that the model took leeway in creating expressions as long as the output looked natural.

5. Ablation Study

As mentioned in Section 3.3, we introduce five types of pseudo errors: rotation, translation, scaling, distortion, and color, at the time of training to emulate the different errors we face during inference. In this section, we perform an ablation to show the effects (at the time of inference) of removing each error during training. In each subsection, we try to remove the errors one at a time. i.e., as we remove rotation, the remaining four errors are still present while training. To showcase the clear distinction between the foreground and the background, we turn off the color error for all the ablations.

As clearly depicted in Fig. 2, each error causes a degradation in the output. The leftmost column in the figure shows the effect of not introducing the color normalization error. This leads to sub-optimal blending between the source and target face with significant artifacts. Similarly, the scale and rotation pseudo errors are also extremely important, as shown in the same figure, Fig. 2. Removing the

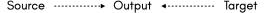




Figure 3: Sample output of blending using the classical technique of Poisson blending.

scaling error causes the blended face to be on a different scale. On the other hand, the rotation error forces the faces to be aligned, making it easier for the algorithm to blend. Finally, without the translation error, the source face does not fit the target face giving rise to an unstructured output. A conjunction of these different errors leads to a setting where the model can blend the given videos spatially and temporally. Affine transformation is a combination of scaling, rotation, and translation. Therefore, removing one of these errors does not confuse the model of the underlying task of alignment. The model still performs the task well and can fit the irregular face shape into the background. However, distortion error (as shown in the figure's last column) is very important. Without the distortion error (which is, in fact, the non-linear transformation), the model struggles to warp the face in a way that best fits the background. This causes the foreground to go out of the background and generate unnatural outputs.

6. Additional Results

6.1. Poisson Blending vs Neural Blending

In this section, we observe that the blending approach fails to produce convincing results by simply applying a heuristic blending technique like Poisson blending on the heuristically aligned frames. The neural blending approach learns a non-linear transformation and blending strategy on the given input that cannot be emulated with a heuristic blending approach like Poisson blending. Poisson blending performs blending well when the source and the target faces are well aligned. It fails to generalize to cases where there is a difference between the source and the target faces, and learning an affine transformation no longer suffices. Faces are rigid bodies, and a rigid-body transformation doesn not suffice for cases with considerable head differences between the source and the target frames.

Moreover, Poisson blending requires precise alignments and masks to paste the source face onto the target face. A sample output of Poisson blending is shown in Fig. 3. The blending was performed after the heuristic alignment step, as shown in Fig. 3 of the main paper. As can be seen, even though the images were blended, the output looks unnatural and distorted.

6.2. Accuracy vs Inference Trade-Off

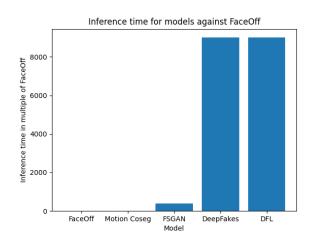


Figure 4: Comparison on the time needed for performing video-to-video face swapping. Faceoff is considered to need $1 \times$ inference time optimization, every other model is plotted relative to FaceOff's inference time. Motion Co-seg: $1.5 \times$, FSGAN: $400 \times$, DeepFakes: $9000 \times$, and DeepFace-Labs: $9000 \times$.

In graph 4, we demonstrate the huge disparity between the inference times of our approach against SOTA approaches DeepFakes, DeepFaceLabs (denoted by DFL), Motion Coseg, and FSGAN. Our approach and motioncoseg are one-shot approaches, and do not require further finetuning. Fsgan provides two modes of inference, a faster inference and an inference that requires finetuning the output. We used the second approach to further improve Fsgan's results and achieved the finetuning in 5 minutes for qualitative results. Quantitative scores were computed without any optimization. Deepfakes and Deepfacelabs require considerable amount of time to achieve reasonable faceswapping and they work on a pair of videos with heavy compute. Even though our approach is one shot, we outperform existing approaches in the SPI metric, as mentioned in the main paper. We achieve the best SPI of 0.38 over all the baseline approaches.

References

- [1] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr. Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, Pingyu Wu, Bo Zhou, and Weiming Zhang. Deepfacelab: Integrated, flexible and extensible face-swapping framework, 2020.
- [2] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2, 2019.