

REFERRING TO THE MAIN ARTICLE, LINE NUMBER IS SPECIFIED

A EXPERIMENTAL SETTINGS

A.1 NETWORK ARCHITECTURE

(Line 204-205: Please refer to Sec. 9 for the details of U-Net network architecture.)

A.1.1 GENERATOR

Our network generator consists of a shared recurrent block, which is essentially a U-Net-based network which comprises both up-sampling and down-sampling layers. For every set of 3 input frames, we mask the frames using input and target ages, thereby increasing the number of input channels from 3 to 5. Each of these input frames has a resolution of 512×512 . In the first layer of the recurrent block, we concatenate three input frames with a 64-channel hidden state and a 3-channel output image from the previous iteration. This combination results in a spatial image with a total of 82 channels. Such a detailed configuration ensures that the spatial image accurately captures both the current input data and the information processed in the preceding steps. Finally, we obtain a 67-channel spatial image that is subsequently split into a delta image and a hidden state. For detailed architectures of these layers and blocks, please refer to Table 4 for down-sampling, Table 5 for up-sampling, Table 6 for the recurrent block, and Table 7 for generator architecture.

Table 4: Down-sampling layer.

Layer	Output
Input	$w \times h \times c$
MaxBlurPool	$w/2 \times h/2 \times c$
3×3 Conv + LeakyReLU	$w/2 \times h/2 \times 2c$
3×3 Conv + LeakyReLU	$w/2 \times h/2 \times 2c$
Output	$w/2 \times h/2 \times 2c$

Table 5: Up-sampling layer.

Layer	Output
Input	$w \times h \times c$
BlurUpSample	$2w \times 2h \times c$
3×3 Conv + LeakyReLU	$2w \times 2h \times c/2$
3×3 Conv + LeakyReLU	$2w \times 2h \times c/2$
Output	$2w \times 2h \times c/2$

A.1.2 DISCRIMINATOR

In our architecture, we use two discriminators. The first discriminator, specifically focused on image quality, is based on PatchGAN Isola et al. (2017). This consists of three downsampling layers followed by two convolution layers. The output frames are concatenated by their target age masks in a channel-wise manner, resulting in a 4-channel input image. Additionally, we process all output frames independently by concatenating them batch-wise as shown in Table 8.

We also incorporate another discriminator with 3D convolutions, referred to as the video discriminator, to evaluate the realism of motion. Our video discriminator consists of three downsampling layers. The input is created by concatenating the target input age mask and the output images corresponding to three consecutive frames, resulting in a four-channel input. The architecture of the video discriminator is illustrated in Table 9. It is noted that both of our discriminators consist of LeakyReLU in every layer, which is not shown in Table 8 and Table 9.

A.2 IMPLEMENTATION SETTINGS

(Line 283-284: We provide the additional details in the supplementary document)

Table 6: Recurrent block.

Layer	Output
Input (Video)	$512 \times 512 \times 3 \times 5$
Reshape	$512 \times 512 \times 15$
Previous Hidden State	$512 \times 512 \times 64$
Previous Output	$512 \times 512 \times 3$
Concatenation	$512 \times 512 \times 82$
3×3 Conv + LeakyReLU	$512 \times 512 \times 64$
3×3 Conv + LeakyReLU	$512 \times 512 \times 64$
DownSampleLayer	$256 \times 256 \times 128$
DownSampleLayer	$128 \times 128 \times 256$
DownSampleLayer	$64 \times 64 \times 512$
DownSampleLayer	$32 \times 32 \times 1024$
UpSampleLayer	$64 \times 64 \times 512$
UpSampleLayer	$128 \times 128 \times 256$
UpSampleLayer	$256 \times 256 \times 128$
UpSampleLayer	$512 \times 512 \times 64$
1×1 Conv	$512 \times 512 \times 67$
Output Delta Image	$512 \times 512 \times 3$
Output Hidden State + LeakyReLU	$512 \times 512 \times 64$

Table 7: Generator architecture. N represents the number of frames in the input sequence.

Layer	Output
Input (Video)	$512 \times 512 \times N \times 3$
Recurrent Blocks ($\times N$)	$512 \times 512 \times N \times 67$
Output (Video)	$512 \times 512 \times N \times 3$

Table 8: Architecture of image discriminator.

Layer	Output
Video with Target Mask	$512 \times 512 \times 4$
4×4 Conv	$256 \times 256 \times 64$
4×4 Conv	$128 \times 128 \times 128$
4×4 Conv	$64 \times 64 \times 256$
4×4 Conv (Stride = 1)	$64 \times 64 \times 512$
4×4 Conv (Stride = 1)	$64 \times 64 \times 1$

Table 9: Architecture of video discriminator.

Layer	Output
Video with Target Mask	$512 \times 512 \times 3 \times 4$
4×4 3D Conv	$256 \times 256 \times 32 \times 4$
4×4 3D Conv	$128 \times 128 \times 64 \times 4$
4×4 3D Conv	$64 \times 64 \times 128 \times 4$
4×4 3D Conv (Stride = 1)	$64 \times 64 \times 256 \times 4$
4×4 3D Conv (Stride = 1)	$64 \times 64 \times 1 \times 4$

In this section, we describe our experimental setup for video re-aging training data. We utilized a total of 4,248 subjects to train our network, generating 14 videos per subject. These subjects were divided into 14 classes, covering a wide age range from 18 to 85. Each video consists of 57 frames for training. We applied a cumulative probability of blur detection (CPBD) threshold of 0.5 to ensure sharpness, reducing the prevalence of blurry videos, especially in those with lower CPBD values. Higher CPBD videos, exhibiting fewer dynamic poses, led us to maintain this threshold to balance sharpness and dynamic representation. The sample images of our generated dataset are shown in Fig. 7. For testing, we selected 20 videos from VFHQ Xie et al. (2022) and 85 videos from CelebV-

HQ Zhu et al. (2022) for each Young \rightarrow Old direction (target ages: 65, 75, 85) and Old \rightarrow Young direction (target ages: 18, 25, 35).

During training, we set input and output ages randomly without imposing any conditions that both ages cannot be equal in the same iteration. This approach enables the reconstruction of the input image when the input and target ages are the same. We used a learning rate of 0.0001 for 250K iterations with a batch size of 4. Additionally, we introduced temporal augmentation in our training. We randomly selected a frame interval Δt from [3, 5, 7]. Reverse augmentation is applied to every frame sequence with a 0.5 probability. We implemented our code in the PyTorch framework and trained our model with a single A100 GPU.

EXPERIMENTAL SETTINGS FOR ABLATION STUDY

For the ablation study on reenactment methods in Table 3 (a), we employed 5 videos of 10 subjects, each with 2 target ages (18 and 85), totaling 100 videos. Table 3 (b), which shows the different interpolation methods, utilizes 21 videos from Xie et al. (2022) for testing. In the experiments exploring various training configurations of OSFV Wang et al. (2021a) (Table 10), we use a relatively large test set comprising 100 videos, as this step is crucial for generating quality data.

B ADDITIONAL QUALITATIVE COMPARISON

In this section, we provide additional comparison results to evaluate our method against state-of-the-art approaches.

Young \rightarrow Old. Fig. 8 shows the comparison results for *Young \rightarrow Old*. The results indicate that SAM Alaluf et al. (2021) emphasizes on target age and does not retrain the image fidelity especially for side-poses. This tendency is also observed for HRFAE Yao et al. (2021b). The diffusion based model DIFF-AE Preechakul et al. (2022) and CUSP Gomez-Trenado et al. (2022) often produce artifacts in output images. Whereas AgeTransGAN Hsu et al. (2022) generate artistic images that appears to be unnatural.

Old \rightarrow Young. We also present our comparison results for *Old \rightarrow Young* task in Fig. 9. These results show the similar tendency in which HRFAE Yao et al. (2021b) and SAM Alaluf et al. (2021) fail to recover the facial details and suffer from significant artifacts. The results indicate that DIFF AE Preechakul et al. (2022) lacks control on target ages and consists various artifacts in their results with variant ages.

C ABLATION OF TRAINING CONFIGURATIONS OF OSFV

(Line 162: (Footnote) Additional details are provided in the supplementary materials.) For generation of key frames, we utilize the unofficial implementation² of OSFV Wang et al. (2021a) trained on VoxCeleb Nagrani et al. (2017) at 256² resolution. We denote this configuration as model ‘A’. We evaluated the performance on 100 Xie et al. (2022)’s test set For the training at 512², we experimented with four different training configurations, as detailed in Table 10. One of the approaches involved a naive upscaling of the images to 512², resulting in a notably poor CPBD. Afterward, we fine-tuned the existing pretrained model on VFHQ Xie et al. (2022) at 512², resulting in improved quality and designated as model ‘B’. We also conducted training from scratch at 256² on the VFHQ dataset, following Wang et al. (2021a). While SSIM and PNSR showed improvement compared to model ‘A’, CPBD remained lower. This model is referred to as model ‘C’. Therefore, we fine-tuned model ‘C’ at 512² and observed an overall improvement in the dataset’s quality, designated as model ‘D’. For the test set, we randomly selected 100 videos from Xie et al. (2022)’s test set. Our configurations are summarized as follows:

A: Publicly available model trained at resolution 256²

B: Fine-tuning model A on VFHQ at resolution 512²

C: Training Wang et al. (2021a) from scratch on VFHQ at resolution 256²

D: Fine-tuning C on VFHQ at resolution 512²

²<https://github.com/zhengkw18/face-vid2vid>

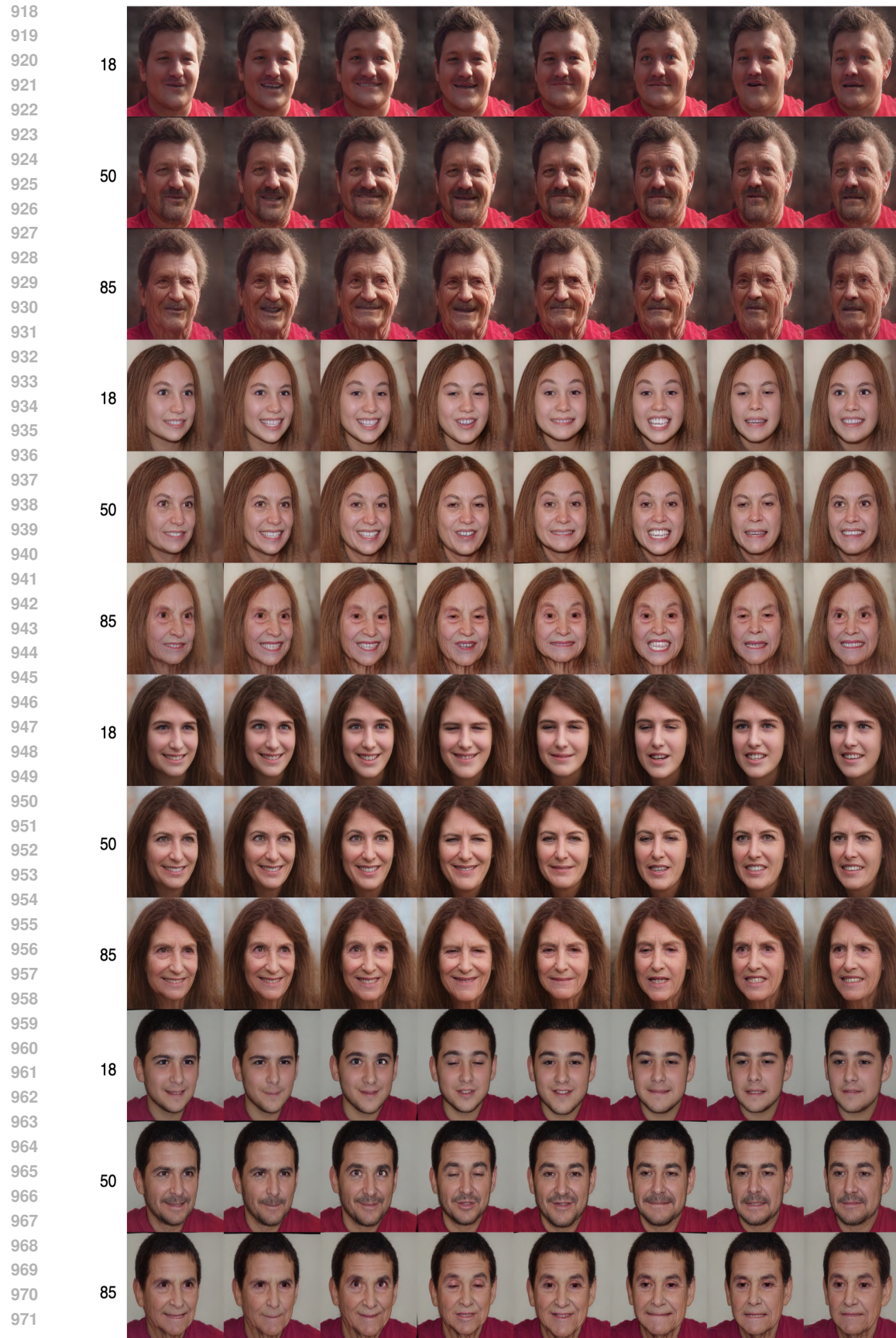


Figure 7: Data samples in our proposed video dataset with ages 18, 50, and 85.



Figure 8: Qualitative comparison with existing state-of-the-art methods on CelebV-HQ dataset. The target age is set to 85.

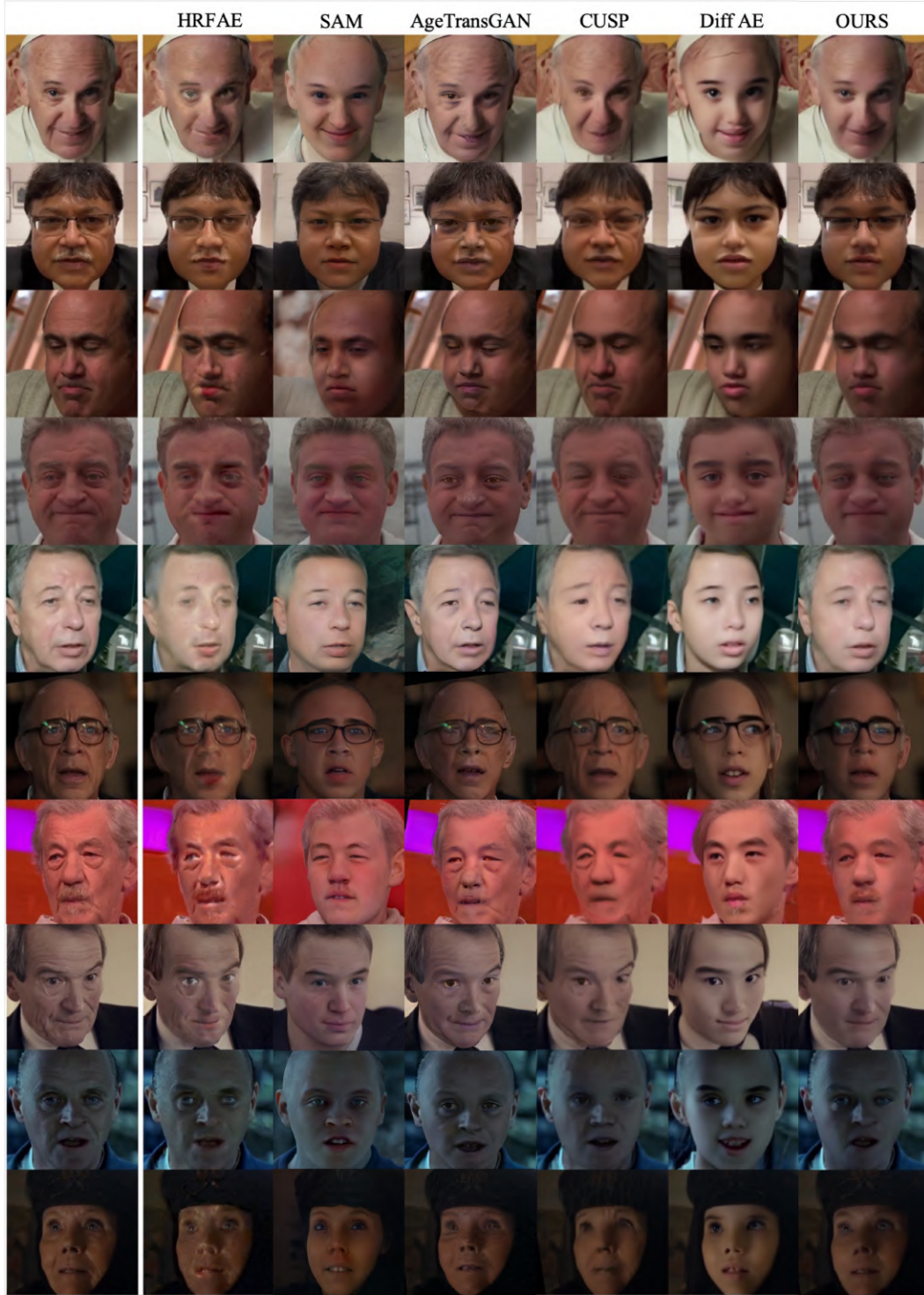


Figure 9: Qualitative comparison with existing state-of-the-art methods on CelebV-HQ dataset. The target age is set to 18.



Figure 10: Limitations of our approach.

Table 10: Ablating performance of OSFV Wang et al. (2021a) with different training configurations.

Method	Resolution	PSNR \uparrow	SSIM \uparrow	CPBD \uparrow
A	256 \times 256	17.553	0.657	0.149
B	512 \times 512	18.476	0.665	0.442
C	512 \times 512	19.200	0.697	0.223
D	512 \times 512	19.519	0.683	0.487

D LIMITATIONS AND FUTURE WORKS

Despite of our work surpasses current state-of-the-art methods, we have observed that our method often struggles to preserve the facial hairs. The limitation of our methods are shown in Fig. 10. Our empirical results found that this problem arises due to SAM Alaluf et al. (2021). Therefore, leveraging alternative methodologies that are closely aligned to our aspirations may improve the performance. Additionally, we employ simple encoder-decoder within recurrent block architecture. One can explore more advanced network architectures, focusing on refining the transformation of facial shapes. This will likely lead to enhancements in age transformation capabilities and temporal consistency. Furthermore, we have utilized Wang et al. (2021a) for face reenactment and Reda et al. (2022) for frame interpolation, integrating models such as Zhang et al. (2023a) could potentially lead to the creation of even more realistic videos and enhancing our performance. These tasks require further investigation by future researchers.

E ETHICAL STATEMENT

Our proposed video dataset comprises synthetic videos which heavily relies on StyleGAN images, trained on FFHQ dataset Karras et al. (2019). We acknowledge the potential biases inherited from StyleGAN and FFHQ. Recognizing the societal threat or risk of misuse of our work, we explicitly disapprove of any malicious applications of our research. Our primary intent is to contribute positively for the production and advertisement industry.