OPHAvatars: One-shot Photo-realistic Head Avatars

Shaoxu Li,1

John Hopcroft Center for Computer Science Shanghai Jiao Tong University Shanghai, China lishaoxu@sjtu.edu.cn

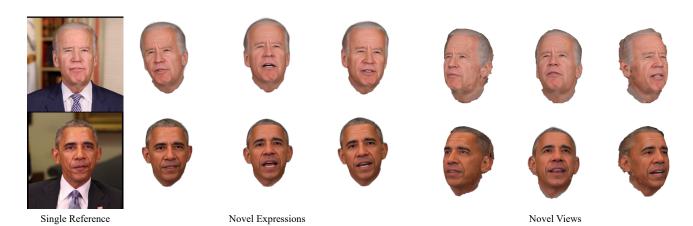


Figure 1: Given a single reference image, our method optimizes a deformable neural radiance field to synthesize photo-realistic animatable 3D neural head avatars. The resulting head avatar can be viewed under novel views and animated with novel expressions.

Abstract

We propose a method for synthesizing photo-realistic digital avatars from only one portrait as the reference. Given a portrait, our method synthesizes a coarse talking head video using driving keypoints features. And with the coarse video, our method synthesizes a coarse talking head avatar with a deforming neural radiance field. With rendered images of the coarse avatar, our method updates the low-quality images with a blind face restoration model. With updated images, we retrain the avatar for higher quality. After several iterations, our method can synthesize a photo-realistic animatable 3D neural head avatar. The motivation of our method is deformable neural radiance field can eliminate the unnatural distortion caused by the image2video method. Our method outperforms state-of-the-art methods in quantitative and qualitative studies on various subjects.

Introduction

Recently, some methods have synthesized photo-realistic avatars leveraging neural radiance fields. As the most popular method for novel view synthesis, NeRF received much attention because of its superb photo-realistic rendering quality. Avatar creation and animation is a prevalent research task with broad application prospects. It's natural to explore the avatar generation with NeRF. Although some worthy

works have been proposed, a one-shot avatar from an image is still underexplored. This work explores the photo-realistic avatar synthesis from a single portrait image.

Earlier works warp the source image with motion fields between the source and driving images. These methods try to simulate the 3D motion with extracted keypoints or other features. Some following works tend to resort to 3D head priors for better performance. Compared with 2D methods, 3D methods can better leverage the human face's unique geometry. In contrast to warp methods, generative methods also perform well, with adequate training images. For face rendering, explicit and implicit methods have strong points and weaknesses.

In this paper, we propose one-shot photo-realistic head avatars(OHPAvatars), which can synthesize a high-quality animatable avatar with only a reference face image. Figure 1 shows two animation examples. Given a single reference image, OHPAvatars can synthesize a high-quality avatar animated with novel expressions and views. Although plenty of works can accomplish dynamic avatar synthesis, the research of one-shot avatars from images remains underexplored. HeadNeRF(Hong et al. 2022b) and OTAvatar(Ma et al. 2023) are similar works to ours. HeadNeRF can not preserve the identity well, and the dynamic rendering is unnatural. OTAvatar needs time-consuming training, and the

rendering quality is unsatisfactory.

Our method builds off video2avatar works with NeRF. Given a short video, a photo-realistic animated avatar can be obtained with a deformable neural radiance field, assisted by 3DMM priors. We employ an image2video method to get through the pipeline of image2avatar. To improve the poor rendering quality caused by inconsistency and low quality of the coarse video, we propose to update the avatar with a blind face restoration method iteratively. After the first avatar optimization with the coarse video, we render images with expressions and poses. We execute blind face restoration on all images and retrain the avatar. After several rendertrain, our method can synthesize a photo-realistic animatable 3D neural head avatar. Figure 2 shows the detail of our pipeline.

In summary, our contributions are listed as follows:

- We propose a one-shot approach for creating a photorealistic animatable 3D neural head avatar with a reference face image.
- We propose eliminating the inconsistency of imagedriven methods with neural radiance field and iteratively updating avatar quality with blind face restoration methods. The pipeline can be transferred to other domains.
- We demonstrate remarkable results of our method through extensive experiments.

Related Work

Talking Head

Talking Head aims to synthesize photo-realistic videos, given source images and driving signals. Driving signals can be image, audio or expression coefficients. FOMM(Siarohin et al. 2019) proposed to extract keypoints mapping between source and driving images, warp the feature of the source image and then decoder for image synthesis. Facevid2vid(Wang, Mallya, and Liu 2021) proposed to estimate motion fields of feature points based on keypoints of source and driving images and warp the feature from source images based on the motion fields. Face2face(Thies et al. 2016) used 3D rendering to maintain the shape and illumination attributes when transferring expression and refining mouth details with a mouth retrieval algorithm. HeadGAN(Doukas, Zafeiriou, and Sharmanska 2021) used additional 3DMM mesh fitting results to assist the dense image flow prediction. DaGAN(Hong et al. 2022a) proposed to capture dense 3D geometric information and estimate sparse facial keypoints for motion estimation. The depth information and motion fields warp the original image. StyleHEAT(Yin et al. 2022) integrated pre-trained StyleGAN(Karras et al. 2020) to generate high-resolution talking face prediction by warping low-resolution feature maps. PIRenderer(Ren et al. 2021) used a subset of 3DMM parameters as motion descriptors, a mapping network maps the input motion descriptors to latent space variables, a warping network generates a rough image, and an editing network generates the final image by editing the rough result.

Some methods pursued audio-driven talking head synthesis. Wav2Lip(Prajwal et al. 2020) constructed the encoder-decoder structure by using audio, pictures synchronized with

audio and pictures not synchronized with audio as input, and realizes the audio-driven mouth video. MakeItTalk(Zhou et al. 2020) used a pre-trained model to extract facial feature points and trained a baseline model of speech-driven feature points. SadTalker(Zhang et al. 2022b) modeled the expression and head pose separately in 3D face parameters and used the Wav2Lip model to complete the mouth synthesis.

Neural Radiance Fields

Neural radiance fields(NeRF)(Mildenhall et al. 2020) is the latest method for novel view synthesis with differentiable volumetric rendering, which has received much attention because of its photo-realistic rendering quality. NeRF uses an MLP to store the 3D scene itself or its features. For avatar synthesis, characteristic features are embedded into the encoding parameters of NeRF. Some methods combine neural radiance fields and generative adversarial networks to complete scene synthesis and editing(Schwarz et al. 2020; Niemeyer and Geiger 2021; Chan et al. 2021, 2022). Portrait manipulation can be accomplished with GAN inversion and latent code adjustment for these GAN-based methods. OTAvatar(Ma et al. 2023) proposed to decouple motion-related and motion-free latent code in inversion optimization by prompting the motion fraction of latent code ahead of the optimization using a decoupling-by-inverting strategy.

Others resort to 3DMM prior to realizing animated neural radiance field avatars. PVA(Raj et al. 2020) enabled portrait generation with a novel parameterization that combines NeRF with local, pixel-aligned features. AD-NeRF(Guo et al. 2021) was the first to propose a speech-driven neural radiance field head model, which completes the speechto-speaker mapping. The method is trained on a speech video of the target person, and neural radiance fields implicitly represent the dynamic human head and upper body torso. HeadNeRF(Hong et al. 2022b) collected and processed three large-scale face image datasets, based on which they designed a nerF-based parameterized head representation. NeRFace(Gafni et al. 2021) can synthesize highquality faces based on controllable pose and expression, given a short face video. NeRFace extract pose and expression with 3DMM and encode them into the input of the neural radiance field. IMAvatar(Zheng et al. 2022) proposed to learn blendshapes and skinning fields to represent the expression-and pose-related deformations of implicit head avatars. NHA(Grassal et al. 2021) obtained high-precision meshes using a shape network and outputs high-quality synthetic images using a texture network. INSTA(Zielonka, Bolkart, and Thies 2023) adopted a similar strategy to IMAvatar, except that INSTA replaced blendshapes and skinning fields with mesh faces extracted by 3DMM and adopted multi-resolution hash coding to accelerate training.

Blind Face Restoration

Blind face restoration aims at recovering high-quality portraits from low-quality face photos. The restoration tasks include face denoising, face deblurring, face super-resolution, face artifact removal and blind face restoration. The restoration solutions include geometric prior-based methods, refer-

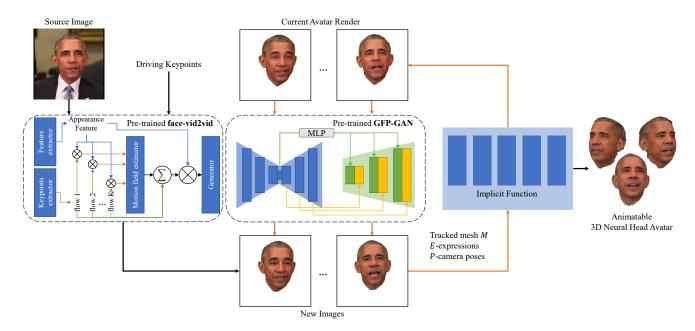


Figure 2: Overview. Our method synthesizes an animatable 3D neural head avatar and gradually updates the avatar by updating head images: (1) a coarse video is generated with a pre-trained face-vid2vid, given a source image and driving keypoints, (2) an animatable 3D neural head avatar is trained with the coarse video, (3) the avatar render images and update images with a pre-trained GFP-GAN, (4) animatable 3D neural head avatar continues training with new images. The iteration of (3)(4) repeats several times for high-quality avatar creation.

ence prior-based methods, generative prior-based methods and non-prior-based methods(Tao et al. 2022). Geometric prior-based restoration methods utilize the unique facial geometry characteristics to restore face images, such as facial landmarks(Chen et al. 2018; Kim et al. 2019), face parsing maps(Chen et al. 2021) and facial heatmaps(Yu et al. 2018). Reference prior-based restoration methods use guidance from additional high-quality face images, such as the facial structure or facial component dictionaries(Li et al. 2018, 2020b,a; Dogan, Gu, and Timofte 2019). Generative models aim to synthesize high-quality images with abundant training images from specific domains. Generative priorbased restoration methods leverage the generative capacity of generative models. For example, PLUSE(Menon et al. 2020), mGANprior(Gu, Shen, and Zhou 2020) restore face by optimizing the latent code from a pre-trained GAN. GFP-GAN(Wang et al. 2021) and GPEN(Yang et al. 2021) use pre-trained GAN as a decoder for face restoration.

Non-prior-based restoration methods try to establish the mapping between low-quality and high-quality face images without any facial priors. For example, BCCNN(Zhou et al. 2015) extracts robust face representations from the low-resolution face image and fuses the representation and the input image for restoration. CBN(Zhu et al. 2016) jointly optimizes FSR and dense correspondence field estimation in a deep bi-network, which benefit each other and recover high-quality face images.

Recently, some methods leveraged attention mechanisms to extract face features, which improve the performance(Chen et al. 2020). And some methods build off

the transformer backbone to build the networks more recently (Wang et al. 2022; Zhang et al. 2022a).

Method

Our method takes as input a human portrait image. As output, our method produces a photo-realistic animatable 3D neural head avatar.

Our method synthesizes a coarse talking head video using driving keypoints features. With the video, we extract parameters and construct a neural radiance field in the canonical space to represent the head and deform the radiance field to learn the avatar with facial expressions. For a photorealistic avatar, we execute blind face restoration with rendered face images and iteratively update the face images for more iterations.

Our method builds off three parts, talking head from one portrait image, blind face restoration, photo-realistic 3D neural head avatar from a video, specifically face-vid2vid(Wang, Mallya, and Liu 2021), GFPGAN(Wang et al. 2021) and INSTA(Zielonka, Bolkart, and Thies 2023).

Background

Face-vid2vid. Face-vid2vid(Wang, Mallya, and Liu 2021) extracts appearance features $F_{appearance}$, 3D canonical keypoints $x_{c,k}$, head pose R_s, t_s and the expression deformations $\delta_{s,k}$ from the source image. Only driving head pose R_d, t_d and the expression deformations $\delta_{d,k}$ are extracted from the driving images. Source keypoints $x_{s,k}$ are calculated with $x_{c,k}$, R_S , t_S and $\delta_{S,K}$. Driving keypoints $x_{d,k}$ are calculated with $x_{c,k}$, R_d , t_d and $\delta_{d,k}$. Face-vid2vid estimate

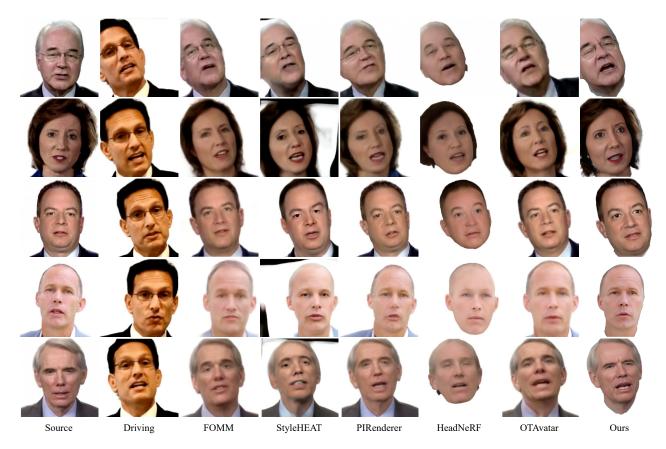


Figure 3: Qualitative result for cross-identity reenactment.

motion flows f_s from the source and driving keypoints for video synthesis. The flow results are combined and fed to the motion field estimation network to produce a flow composition mask m. A linear combination of m and motion flows f_s produces the composited flow field w, which warp the appearance features $F_{appearance}$. Finally, a generator converts the warped feature $w(F_{appearance})$ to the output image. For more details, we direct the reader to the original paper(Wang, Mallya, and Liu 2021).

GFPGAN. GFPGAN(Wang et al. 2021) removes the image degradation and extracts features F_{latent} and $F_{spatial}$ with a degradation removal module, which is a U-Net structure.

$$F_{latent}, F_{spatial} = \text{U-Net}(\mathbf{x}),$$
 (1)

GFPGAN uses the latent features F_{latent} to map the input image to the closest latent codes W in StyleGAN2(). The latent codes W generate the GAN features F_{GAN} .

$$W = MLP(F_{latent}),$$
 (2)

$$F_{GAN} = \text{StyleGAN}(W),$$
 (3)

GFPGAN uses the multi-resolution spatial features $F_{spatial}$ to modulate the StyleGAN2 features. At each resolution scale, a pair of affine transformation parameters α, β are generated from spatial features $F_{spatial}$. And the modulation is carried out through channel-split spatial feature transform

(CSSFT) layers. For more details, we direct the reader to the original paper(Wang et al. 2021).

$$\alpha, \beta = \text{Conv}(F_{spatial}),$$
 (4)

$$F_{output} = \text{SC-SFT}(F_{GAN}|\alpha,\beta),$$
 (5)

INSTA. Neural radiance field(Mildenhall et al. 2020) based head synthesis has attracted much attention for its high quality. INSTA(Zielonka, Bolkart, and Thies 2023) is a neural radiance field-based method for photo-realistic digital avatars reconstructing. In practice, INSTA is trained on a single monocular RGB portrait video and can reconstruct a digital avatar that extrapolates to unseen expressions and poses. For a given video, inputs for avatar optimization include head images I_i , the intrinsic camera parameters $K \in \mathbb{R}^{3\times 3}$, tracked FLAME(Li et al. 2017) meshes $\{M_i\}$, facial expression coefficient $\{E_i\}$, and head poses $\{P_i\}$. IN-STA constructs a neural radiance field in the canonical space to represent the head. A mapping function $\phi(p, M_i)$ projects the points from time-varying deformed space to the canonical space, leveraging the tracked FLAME(Li et al. 2017) meshes $\{M_i\}$ and a predefined mesh in canonical space $\{M_canon\}$. Differentiable volumetric rendering is used to optimize the radiance field. The animatable dynamic neural radiance field is implemented using the Nvidia NGP C++ framework(). In practice, INSTA samples 1700 frames from the video and trains the network for 32k optimization

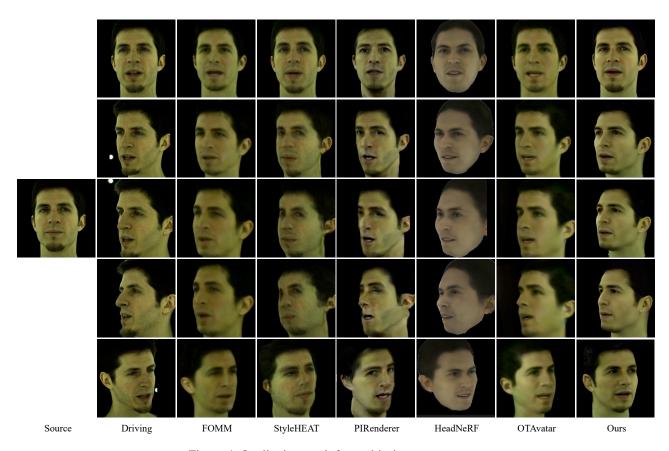


Figure 4: Qualitative result for multi-view reenactment.

steps. The resulting head avatar can be viewed under novel views and expressions with the optimized deformable radiance field. For more details, we direct the reader to the original paper(Zielonka, Bolkart, and Thies 2023).

One-shot Photo-realistic Head Avatars

A short monocular RGB video can reconstruct neural radiance field-based photo-realistic digital avatars. Our method works by generating a coarse video with a pre-trained image2video model and embedding the data enhancement into the training of the neural radiance field by updating the rendered images with a blind face restoration model.

In this section, we first illustrate the process of generating a coarse video, then describe the iterative dataset update strategy, and finally, discuss our method's implementation details. The following mentioned head images are segmented from original videos, which do not include background and body.

Coarse Avatar Generation. We use face-vid2vid(Wang, Mallya, and Liu 2021) to obtain a coarse video. As a talking head synthesis method, it inputs a source image I_s and a sequence of unsupervised-learned 3D keypoints $x_{d,k}$. The output coarse talking head video contains a set of portrait images with different expressions and poses, which provide abundant training information for avatar synthesis. With these initial portrait images, we segment the background and

foreground to obtain training images $\{I_{d,i}^{coarse}\}$. FLAME meshes and parameters are calculated for later avatar optimization. We use INSTA(Zielonka, Bolkart, and Thies 2023) to train an animatable 3D neural head avatar, a deformable neural radiance field with 3DMMs parameters as reference.

With the increase of view angle, the synthesis quality of the coarse video decreases. For avatar synthesis, large view angle images can improve the quality of the avatar. Fortunately, as a 3D representation, NeRF can maintain the stability of the scene geometry. In this case, the stable facial geometry promises the stability of the identity. For the deformable neural radiance field, the coarse video's distortions only impact the avatar's visual quality but not the geometry. But we cannot obtain high-quality avatars with low-quality images from the coarse video.

Iterative Dataset Update. Even though NeRF is a 3D representation method that ensures high-quality image synthesis, our avatar is of low quality due to the initial video's poor image quality and inconsistency. For high-quality synthesis, we propose an iterative dataset update with blind face restoration. After acquiring the coarse avatar, we render the avatar into images $\{I_{d,i}^{render}\}$, conditioned with various expressions and views. Then a face image restoration is executed on all the images, specifically GFPGAN(Wang et al. 2021). With new images and rendered parameters, we train

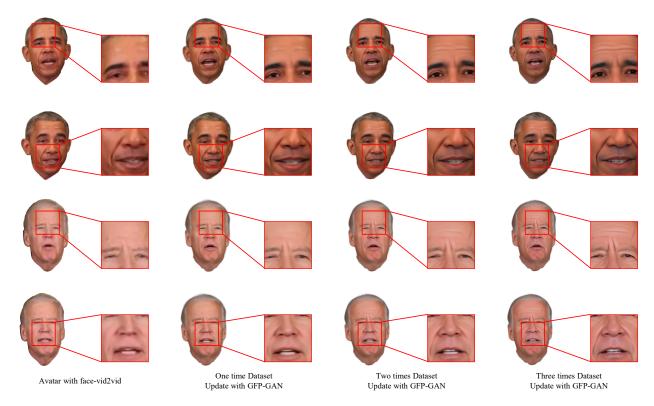


Figure 5: Ablation Study: We compare face-vid2vid and one/twp/three times dataset update avatar synthesis results with GFP-GAN.

the avatar again. The process above can be conducted repeatedly. The rendering quality of the avatar increase with the iteration. After several iterations, we can obtain a photorealistic animatable 3D neural head avatar.



Figure 6: The face restoration results with the iterations.

Although the rendering quality increases with iteration, the results may not be as expected. For a coarse image, a restoration tries to fill in details. A well-restored image may be inconsistent with the original image, although clear in detail. Figure 6 shows the restoration results of a portrait image. The image becomes sharper with the iterations, but too many details are added.

Results

Experiment Setting

In this section, we first describe the implementation details of our method. Then we compare our method with state-ofart face reenactment and avatar creation methods. At last, we execute an ablation study to illustrate the effectiveness of our designs.

Implementation details. Abundant poses and expressions are needed for high-quality avatar training. We create around 2-3min videos with manually driving keypoints or driving videos for coarse video generation. The coarse video is subsampled to 25fps and resized to 512×512 resolution. We preprocess the video with the pipeline in INSTA(Zielonka, Bolkart, and Thies 2023). Firstly, we use background foreground segmentation using robust matting(Lin et al. 2021) and an off-the-shelf face parsing framework(Yu et al. 2020) for image segmentation and clothes removal. Secondly, we use MICA(Zie 2022) to obtain optimized FLAME mesh, expression parameters, and pinhole camera intrinsic and extrinsic parameters. With these parameters and head images, we use INSTA to optimize an avatar. Our model is implemented in Pytorch using one RTX 3090. For each avatar optimization, the model is trained for 30000 iterations. Thanks to the fast training of INSTA with multiresolution encoding, each iteration can be accomplished in less than 20 minutes.

Dataset. We conduct cross-identity reenactment experiments on HDTF(Zhang et al. 2021) dataset and multi-view reenactment experiments on Multiface(hsin Wuu et al. 2023) dataset. And the quantitative comparison is conducted on the same dataset. For the ablation study, we use two celebrity photos for illustration.

Table 1: Quantitative comparisons on multi-View reenactment and cross-identity reenactment.

	Multi-View Reenactment								Cross-Identity Reenactment				
	PSNR↑	SSIM↑	CSIM↑	AED↓	APD↓	AKD↓	LIPIPS↓	FID↓	CSIM↑	AED↓	APD↓	AKD↓	FID↓
FOMM	20.75	0.639	0.505	2.004	0.545	5.052	0.308	101.6	0.672	3.196	0.500	4.198	113.2
PIRenderer	20.04	0.586	0.493	2.203	0.680	6.566	0.299	100.6	0.632	3.018	0.498	4.977	103.7
StyleHEAT	20.03	0.632	0.387	2.179	0.472	5.522	0.284	123.8	0.614	2.860	0.471	3.592	239.1
HeadNeRF	17.60	0.546	0.239	2.086	0.776	4.166	0.367	212.3	0.282	2.873	0.567	3.465	233.0
OTAvatar	21.19	0.657	0.574	1.874	0.428	3.731	0.288	137.3	0.694	2.850	0.405	4.307	101.8
Ours	21.68	0.690	0.634	1.495	0.265	4.135	0.231	95.3	0.503	1.439	0.270	4.245	97.4

Baseline methods. We compare our method with state-of-the-art 2D and 3D methods. For image-driven method, FOMM(Siarohin et al. 2019) and face-vid2vid(Wang, Mallya, and Liu 2021) achieve SOTA performance. For 2D methods with 3DMM coefficients, StyleHEAT(Yin et al. 2022) and PIRenderer(Ren et al. 2021) are the SOTA methods. For 3D methods, HeadNeRF(Hong et al. 2022b) and OTAvatar(Ma et al. 2023) achieve SOTA perfomance. As a component of our method, face-vid2vid(Wang, Mallya, and Liu 2021) is compared in the ablation study.

Evaluation metrics. For comprehensive comparative analysis, we adopt peak-signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and learned perceptual image patch similarity (LPIPS)(Zhang et al. 2018) for visual quality, frechet inception distance (FID)(Heusel et al. 2017) and cosine similarity of identity embedding (CSIM)(Deng et al. 2022) for image realism. And inspired by OTAvatar(Ma et al. 2023), we adopt average expression distance (AED), average pose distance (APD), and average keypoint distance (AKD) for facial expression and pose transfer.

Evaluation Result

Qualitative comparison. For cross-identity reenactment, we evaluate the transfer quality of the synthesis results on HDTF(Zhang et al. 2021) dataset. Figure 3 shows the qualitative results. Compared with warping-based methods FOMM, StyleHEAT and PIRender, our methods can maintain the identity better and produce consistent facial geometry. Compared with 3D HeadNeRF, our method reconstruct the identity of the source image better. Compared with OTAvatar, our method ensures higher visual quality. Some pose and expression biases are caused by inaccurate poses and expressions extraction from driving images. And some facial details vary from the source images since our avatar is optimized with the blind face restoration method.

For multi-view reenactment, we evaluate the consistency of the synthesis results in different views on Multiface(hsin Wuu et al. 2023) dataset. Figure 4 shows the qualitative results. Compared with cross-identity reenactment, multi-view reenactment shows each method's large view angel face reenactment ability. StyleHEAT and PIRender can hardly maintain the face geometry. The face identity of HeadNeRF is far from the source image. FOMM and OTAvatar can maintain the face geometry well, while the rendering quality is unsatisfactory. By our method, the avatar can preserve the identity well, with the best rendering quality.

Quantitative comparison. Table 1 shows the quantitative comparison results on multi-view reenactment and cross-identity reenactment. Multi-View Reenactment is computed on Multifaces datasets, and Cross-Identity Reenactment is computed on HDTF datasets. Our method outperforms other methods in terms of most of the criteria.

Ablation Study

We validate the architecture of our method by ablation study. Figure 5 shows the avatar animation results with coarse video from face-vid2vid and one/two/three times dataset update with GFP-GAN. The two celebrity photos in Figure 1 are the source images. The driving poses are extracted from a short video. The dataset update makes rendering images sharper, and the rough facial structures remain unchanged. But the sharp image may be inconsistent with the source image. We suggest using dataset update two times, weighing the source identity and the avatar quality.

Limitations

Our method uses blind face restoration to improve the avatar's quality, which leads to many facial details. Wrinkles from the original portrait are sharpened. More details may deviate from the original identity. Although our method can synthesize novel view images for unseen poses, it's hard to explore extreme novel views, for example, 90°. With the increase of view angle, the rendering quality decreases. With the increase of view angles, the expression coefficients extracted may be inaccurate, which leads to inaccurate rendering expressions.

Conclusion

In this work, we present a novel framework of one-shot photo-realistic head avatars, namely OPHAvatars, using a deformable neural radiance field. The pipeline includes two stages, coarse video generation and neural avatar synthesis. Coarse video generation synthesizes a talking head video with pre-trained image2video methods. Neural avatar synthesis takes the coarse video as input and synthesizes an animatable avatar with a deformable neural radiance field. Then we take a dataset update strategy to improve the performance with a blind face restoration method. After several restoration and training, our method optimizes a photorealistic animatable 3D neural head avatar. Comprehensive experiments were conducted to prove the superiority of our proposed framework.

References

- 2022. Towards Metrical Reconstruction of Human Faces.
- Chan, E.; Monteiro, M.; Kellnhofer, P.; Wu, J.; and Wetzstein, G. 2021. pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis. In *CVPR*.
- Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; Mello, S. D.; Gallo, O.; Guibas, L.; Tremblay, J.; Khamis, S.; Karras, T.; and Wetzstein, G. 2022. Efficient Geometry-aware 3D Generative Adversarial Networks. In *CVPR*.
- Chen, C.; Gong, D.; Wang, H.; Li, Z.; and Wong, K.-Y. K. 2020. Learning Spatial Attention for Face Super-Resolution. In *IEEE Transactions on Image Processing (TIP)*.
- Chen, C.; Li, X.; Lingbo, Y.; Lin, X.; Zhang, L.; and Wong, K. 2021. Progressive Semantic-Aware Style Transformation for Blind Face Restoration. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Chen, Y.; Tai, Y.; Liu, X.; Shen, C.; and Yang, J. 2018. FS-RNet: End-to-End Learning Face Super-Resolution with Facial Priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Deng, J.; Guo, J.; Yang, J.; Xue, N.; Kotsia, I.; and Zafeiriou, S. 2022. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 5962–5979.
- Dogan, B.; Gu, S.; and Timofte, R. 2019. Exemplar Guided Face Image Super-Resolution Without Facial Landmarks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.*
- Doukas, M. C.; Zafeiriou, S.; and Sharmanska, V. 2021. HeadGAN: One-shot Neural Head Synthesis and Editing. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Gafni, G.; Thies, J.; Zollhöfer, M.; and Nießner, M. 2021. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8649–8658.
- Grassal, P.-W.; Prinzler, M.; Leistner, T.; Rother, C.; Nießner, M.; and Thies, J. 2021. Neural Head Avatars from Monocular RGB Videos. *arXiv preprint arXiv:2112.01554*. Gu, J.; Shen, Y.; and Zhou, B. 2020. Image Processing Using Multi-Code GAN Prior. In *CVPR*.
- Guo, Y.; Chen, K.; Liang, S.; Liu, Y.; Bao, H.; and Zhang, J. 2021. AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30.
- Hong, F.-T.; Zhang, L.; Shen, L.; and Xu, D. 2022a. Depth-Aware Generative Adversarial Network for Talking Head Video Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

- Hong, Y.; Peng, B.; Xiao, H.; Liu, L.; and Zhang, J. 2022b. HeadNeRF: A Real-time NeRF-based Parametric Head Model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- hsin Wuu, C.; Zheng, N.; Ardisson, S.; Bali, R.; Belko, D.; Brockmeyer, E.; Evans, L.; Godisart, T.; Ha, H.; Huang, X.; Hypes, A.; Koska, T.; Krenn, S.; Lombardi, S.; Luo, X.; McPhail, K.; Millerschoen, L.; Perdoch, M.; Pitts, M.; Richard, A.; Saragih, J.; Saragih, J.; Shiratori, T.; Simon, T.; Stewart, M.; Trimble, A.; Weng, X.; Whitewolf, D.; Wu, C.; Yu, S.-I.; and Sheikh, Y. 2023. Multiface: A Dataset for Neural Face Rendering. arXiv:2207.11243.
- Karras, T.; Aittala, M.; Hellsten, J.; Laine, S.; Lehtinen, J.; and Aila, T. 2020. Training Generative Adversarial Networks with Limited Data. In *Proc. NeurIPS*.
- Kim, D.; Kim, M.; Kwon, G.; and Kim, D.-S. 2019. Progressive Face Super-Resolution via Attention to Facial Landmark. arXiv:1908.08239.
- Li, T.; Bolkart, T.; Black, M. J.; Li, H.; and Romero, J. 2017. Learning a Model of Facial Shape and Expression from 4D Scans. *ACM Trans. Graph.*, 36(6).
- Li, X.; Chen, C.; Zhou, S.; Lin, X.; Zuo, W.; and Zhang, L. 2020a. Blind Face Restoration via Deep Multi-scale Component Dictionaries. In *ECCV*.
- Li, X.; Li, W.; Ren, D.; Zhang, H.; Wang, M.; and Zuo, W. 2020b. Enhanced Blind Face Restoration with Multi-Exemplar Images and Adaptive Spatial Feature Fusion. In *CVPR*.
- Li, X.; Liu, M.; Ye, Y.; Zuo, W.; Lin, L.; and Yang, R. 2018. Learning Warped Guidance for Blind Face Restoration. In *The European Conference on Computer Vision (ECCV)*.
- Lin, S.; Yang, L.; Saleemi, I.; and Sengupta, S. 2021. Robust High-Resolution Video Matting with Temporal Guidance. arXiv:2108.11515.
- Ma, Z.; Zhu, X.; Qi, G.; Lei, Z.; and Zhang, L. 2023. OTA-vatar: One-shot Talking Face Avatar with Controllable Triplane Rendering. *arXiv preprint arXiv:2303.14662*.
- Menon, S.; Damian, A.; Hu, S.; Ravi, N.; and Rudin, C. 2020. PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. arXiv:2003.03808.
- Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*
- Niemeyer, M.; and Geiger, A. 2021. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Prajwal, K. R.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. 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*, 484–492.
- Raj, A.; Zollhoefer, M.; Simon, T.; Saragih, J.; Saito, S.; Hays, J.; and Lombardi, S. 2020. PVA: Pixel-aligned Volumetric Avatars. In *arXiv*:2101.02697.

- Ren, Y.; Li, G.; Chen, Y.; Li, T. H.; and Liu, S. 2021. PIRenderer: Controllable Portrait Image Generation via Semantic Neural Rendering. arXiv:2109.08379.
- Schwarz, K.; Liao, Y.; Niemeyer, M.; and Geiger, A. 2020. GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019. First Order Motion Model for Image Animation. In *Conference on Neural Information Processing Systems (NeurIPS)*.
- Tao, W.; kaihao, Z.; Xuanxi, C.; Wenhan, L.; Jiankang, D.; Tong, L.; Xiaochun, C.; Wei, L.; Hongdong, L.; and Stefanos, Z. 2022. A Survey of Deep Face Restoration: Denoise, Super-Resolution, Deblur, Artifact Removal. *arXiv* preprint arXiv:2211.02831.
- Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; and Niessner, M. 2016. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2387–2395. Los Alamitos, CA, USA.
- Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021. One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Wang, X.; Li, Y.; Zhang, H.; and Shan, Y. 2021. Towards Real-World Blind Face Restoration with Generative Facial Prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Wang, Z.; Zhang, J.; Chen, R.; Wang, W.; and Luo, P. 2022. RestoreFormer: High-Quality Blind Face Restoration from Undegraded Key-Value Pairs. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yang, T.; Ren, P.; Xie, X.; and Zhang, L. 2021. GAN Prior Embedded Network for Blind Face Restoration in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yin, F.; Zhang, Y.; Cun, X.; Cao, M.; Fan, Y.; Wang, X.; Bai, Q.; Wu, B.; Wang, J.; and Yang, Y. 2022. StyleHEAT: One-Shot High-Resolution Editable Talking Face Generation via Pre-trained StyleGAN. *arxiv*:2203.04036.
- Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; and Sang, N. 2020. BiSeNet V2: Bilateral Network with Guided Aggregation for Real-time Semantic Segmentation. arXiv:2004.02147.
- Yu, X.; Fernando, B.; Ghanem, B.; Porikli, F.; and Hartley, R. 2018. Face Super-Resolution Guided by Facial Component Heatmaps. In *Computer Vision ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part IX*, 219–235.
- Zhang, P.; Zhang, K.; Luo, W.; Li, C.; and Wang, G. 2022a. Blind Face Restoration: Benchmark Datasets and a Baseline Model. arXiv:2206.03697.
- Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

- Zhang, W.; Cun, X.; Wang, X.; Zhang, Y.; Shen, X.; Guo, Y.; Shan, Y.; and Wang, F. 2022b. SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. *arXiv* preprint *arXiv*:2211.12194.
- Zhang, Z.; Li, L.; Ding, Y.; and Fan, C. 2021. 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*, 3661–3670.
- Zheng, Y.; Abrevaya, V. F.; Bühler, M. C.; Chen, X.; Black, M. J.; and Hilliges, O. 2022. I M Avatar: Implicit Morphable Head Avatars from Videos. In *Computer Vision and Pattern Recognition (CVPR)*.
- Zhou, E.; Fan, H.; Cao, Z.; Jiang, Y.; and Yin, Q. 2015. Learning Face Hallucination in the Wild. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 3871–3877.
- Zhou, Y.; Han, X.; Shechtman, E.; Echevarria, J.; Kalogerakis, E.; and Li, D. 2020. MakeItTalk: Speaker-Aware Talking-Head Animation. *ACM Transactions on Graphics*, 39(6).
- Zhu, S.; Liu, S.; Loy, C. C.; and Tang, X. 2016. Deep Cascaded Bi-Network for Face Hallucination. In *European Conference on Computer Vision*.
- Zielonka, W.; Bolkart, T.; and Thies, J. 2023. Instant Volumetric Head Avatars. In *CVPR*.