

LatentAvatar: Learning Latent Expression Code for Expressive Neural Head Avatar

Yuelang Xu
Tsinghua University
Beijing, China
xll20@mails.tsinghua.edu.cn

Hongwen Zhang
Tsinghua University
Beijing, China
zhanghongwen@mail.tsinghua.edu.cn

Lizhen Wang
Tsinghua University
NNKosmos Technology
Beijing, China
wlz18@mails.tsinghua.edu.cn

Xiaochen Zhao
Tsinghua University
NNKosmos Technology
Beijing, China
zhaoxc19@mails.tsinghua.edu.cn

Han Huang
OPPO Research Institute
Beijing, China
huang.h92@outlook.com

Guojun Qi
OPPO Research Seattle
Westlake University
Seattle, United States of America
guojunjq@gmail.com

Yebin Liu
Tsinghua University
Beijing, China
liuyebin@mail.tsinghua.edu.cn



Figure 1: We propose LatentAvatar, an expressive neural head avatar driven by latent expression codes. LatentAvatar is able to capture subtle expressions such as pouting (left) and perform expressive reenactment (right) between different subjects.

ABSTRACT

Existing approaches to animatable NeRF-based head avatars are either built upon face templates or use the expression coefficients of templates as the driving signal. Despite the promising progress, their performances are heavily bound by the expression power and the tracking accuracy of the templates. In this work, we present LatentAvatar, an expressive neural head avatar driven by latent expression codes. Such latent expression codes are learned in an end-to-end and self-supervised manner without templates, enabling our method to get rid of expression and tracking issues. To achieve this, we leverage a latent head NeRF to learn the person-specific latent expression codes from a monocular portrait video, and further design a Y-shaped network to learn the shared latent expression codes of different subjects for cross-identity reenactment. By optimizing the photometric reconstruction objectives in NeRF, the latent expression codes are learned to be 3D-aware while faithfully

capturing the high-frequency detailed expressions. Moreover, by learning a mapping between the latent expression code learned in shared and person-specific settings, LatentAvatar is able to perform expressive reenactment between different subjects. Experimental results show that our LatentAvatar is able to capture challenging expressions and the subtle movement of teeth and even eyeballs, which outperforms previous state-of-the-art solutions in both quantitative and qualitative comparisons. Project page: <https://www.liyebin.com/latentavatar>.

CCS CONCEPTS

- Computing methodologies → Animation; Volumetric models; Motion processing.

KEYWORDS

Facial Reenactment, Expression Transfer

ACM Reference Format:

Yuelang Xu, Hongwen Zhang, Lizhen Wang, Xiaochen Zhao, Han Huang, Guojun Qi, and Yebin Liu. 2023. LatentAvatar: Learning Latent Expression Code for Expressive Neural Head Avatar. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings (SIGGRAPH '23 Conference Proceedings)*, August 6–10, 2023, Los Angeles, CA, USA. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3588432.3591545>

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGGRAPH '23 Conference Proceedings, August 6–10, 2023, Los Angeles, CA, USA

© 2023 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0159-7/23/08.

<https://doi.org/10.1145/3588432.3591545>

1 INTRODUCTION

Creating a 3D head avatar from a monocular video has a great application prospect in digital human, CG Filmmaking, VR and AR, etc. This field has attracted growing attention in recent years. By leveraging the face template prior [Geric et al. 2017; Li et al. 2017] or the implicit field representation [Mildenhall et al. 2020; Park et al. 2019], recent works [Gafni et al. 2021; Gao et al. 2022; Grassal et al. 2022; Xu et al. 2023; Zheng et al. 2022] can recover photo-realistic 3D head avatars using a monocular video. However, despite the promising progress, efficient and expressive control of the head avatar remains unsolved in previous approaches.

When modeling a head avatar, existing methods typically leverage explicit mesh templates [Grassal et al. 2022; Khakhulin et al. 2022; Zheng et al. 2022] or neural implicit representations [Athar et al. 2023; Gafni et al. 2021; Gao et al. 2022; Xu et al. 2023]. Despite the efficiency of face templates, the expression representation power is bound by the linear expression blendshapes or the linear skinning of the face model, which leads to coarse control of avatars and the lack of person-specific detailed expressions. On the other hand, neural implicit representations bypass the constraint of explicit templates and synthesize the facial images directly. However, modeling dynamic 3D heads remain challenging for implicit representations. To control the head avatar, these methods typically resort to conditioning the implicit field with additional expression information, such as the expression coefficients of 3DMM [Geric et al. 2017], FLAME [Li et al. 2017], or FaceWarehouse [Cao et al. 2014b]. Although the avatars created by these methods are not confined by the template topology, the linear expression coefficients of the face templates make it difficult to model high-frequency and detailed person-specific expressions. Moreover, when using the face templates, the misalignment of the tracking results introduces additional deviation in the expression condition. Meanwhile, the expression coefficients and identity coefficients of the face template are easy to be coupled with each other, leading to unexpected artifacts during the cross-identity reenactment of avatars.

To overcome these limitations, we propose LatentAvatar, an expressive neural head avatar driven by latent expression codes. The core idea of LatentAvatar is to learn a latent expression code as the drive signal to animate the head avatar expressively. To achieve this, we first learn a latent head NeRF, where a person-specific latent expression code is learned to drive a customized head NeRF using a monocular portrait video. In the latent head NeRF, the expression code is learned in a self-supervised manner by the photometric reconstruction loss. Such a latent expression code can faithfully capture those high-frequency detailed expressions of the target subject. Besides, by driving the head radiance field, the expression code is also 3D-aware and enables the modeling of viewpoint-consistent avatars. Compared with previous solutions, our method gets rid of the template tracking and expression issues as the latent expression code is learned in a full end-to-end manner without templates.

For cross-identity reenactment, LatentAvatar further leverages a shared latent expression code for the modeling of the shared expression between different subjects. To this end, we introduce a Y-shaped network architecture [Naruniec et al. 2020; Yan et al. 2018] consisting one single shared encoder and two individual decoders.

The shared encoder takes both the avatar and actor images as input to learn a shared latent expression code, which is decoded by two decoders for the individual reconstruction of the input subjects. Finally, a mapping MLP is used to build a bridge between two latent spaces by mapping the shared expression codes to the person-specific one. In this way, LatentAvatar leverages the latent expression code learned in the shared and person-specific settings and enables expressive reenactment between different subjects.

In summary, the contributions of this work can be listed as:

- We propose LatentAvatar, an expressive neural head avatar driven by latent expression codes. Such an expression code is learned in an end-to-end and self-supervised manner without templates, enabling our method to get rid of expression and tracking issues.
- We leverage a latent head NeRF to learn the person-specific latent expression code from a monocular portrait video. The latent expression code is learned to drive the head NeRF, making it 3D-aware while faithfully capturing the high-frequency detailed expressions.
- We further leverage a Y-shaped network to learn a shared latent expression code of different subjects to enable cross-identity reenactment. By bridging the latent expression code learned in shared and person-specific settings, LatentAvatar is able to perform expressive reenactment between different subjects and surpass the performances of previous 2D and 3D solutions.

2 RELATED WORKS

2.1 Head Avatar Modeling

Reconstructing 3D head avatars from monocular videos is an appealing yet challenging task. In the past years, mainstream methods [Cao et al. 2015, 2016; Deng et al. 2019; Hu et al. 2017; Ichim et al. 2015; Nagano et al. 2018] reconstruct mesh-based head avatars based on the morphable face templates tracked in the training portrait video. To handle highly non-rigid contents such as hair, gazes, and teeth, recent methods [Grassal et al. 2022; Khakhulin et al. 2022] leverage neural networks to learn head avatars with dynamic texture and geometry upon the FLAME mesh model [Li et al. 2017]. However, these methods often result in blurred textures due to geometric inaccuracy. To leverage more flexible representations such as the neural implicit fields [Mescheder et al. 2019; Park et al. 2019], IMAvatar [Zheng et al. 2022] proposes to learn head avatars with implicit geometry and texture model [Li et al. 2017], thus gets rid of the topology limitation of the mesh templates. IMAvatar is further extended in PointAvatar [Zheng et al. 2023a] to combine the explicit point cloud with the implicit representation to improve the quality of the rendered images. Since the emergence of NeRF representations [Mildenhall et al. 2020; Park et al. 2021a,b], there are several attempts to exploit its rendering power for neural head modeling [Athar et al. 2023, 2022; Gafni et al. 2021; Guo et al. 2021; Liu et al. 2022; Wang et al. 2022; Zheng et al. 2023b]. Recently, FDNeRF [Zhang et al. 2022] then combine the few-shot NeRF methods [Chen et al. 2021; Yu et al. 2021] to extend these NeRF-based head avatar methods to few-shot reconstruction. For faster avatar modeling, recent methods [Gao et al. 2022; Xu et al. 2023; Zielonka et al. 2022] introduce voxel representation [Fang et al. 2022; Müller

et al. 2022] to solve the problem of low training speed of the NeRF model. However, most of the above solutions rely on the tracked faces templates as they typically use the expression coefficients of templates as the drive signal for avatar animation. Recent methods [Chan et al. 2022; Hong et al. 2022; Sun et al. 2022a, 2023, 2022b; Yenamandra et al. 2021; Zhuang et al. 2022] use large-scale face datasets to train an implicit head template. But it is not easy for these methods to achieve reenactment between different subjects.

There are also research efforts devoted to creating high-fidelity mesh-based head avatars from single-person multi-view synchronized video data [Bi et al. 2021; Lombardi et al. 2018, 2019; Ma et al. 2021; Wang et al. 2021a]. To capture expression information, encoder-decoder networks are typically leveraged to learn latent codes as the compact representations of inputs [Cao et al. 2021; Chu et al. 2020; Lombardi et al. 2019]. Note that these latent codes are person-specific and only used as the self-driven signal for avatar control. To achieve high-resolution rendering, MVP representation [Lombardi et al. 2021] constructs a mixture of volumetric primitives on a coarse mesh and synthesizes high-fidelity images through volumetric rendering in a similar way to NeRF [Mildenhall et al. 2020]. Based on the MVP representation, Cao et al. [Cao et al. 2022] train a generalized head model on a super large-scale multi-view multi-person video dataset. Despite the high-quality results, these methods heavily rely on large-scale multi-view data and even require depth cameras for accurate tracking of the mesh model.

2.2 Facial Reenactment

Existing facial reenactment methods can be roughly divided into three categories: template-based, warping-based, and mapping-based methods. Early works [Cao et al. 2014a, 2013; Li et al. 2012; Thies et al. 2015, 2016; Vlasic et al. 2005; Weise et al. 2011] are typically template-based [Dale et al. 2011; Nirkin et al. 2018; Olszewski et al. 2017] and require a source video for training, which make full use of face priors to fit the target subject with a common morphable template. Recent template-based works [Doukas et al. 2020; Kim et al. 2018; Koujan et al. 2020; Wang et al. 2023; Zakharov et al. 2019] use image-to-image generation networks [Goodfellow et al. 2014] to synthesize photo-realistic images with the template guidance [Gerig et al. 2017]. These methods are extended to few-shot input setting, and further the guidance of template are replaced by semantic maps and facial landmarks [Chen et al. 2020; Korshunova et al. 2017; Natsume et al. 2018; Nirkin et al. 2019a,b; Perov et al. 2021; Zakharov et al. 2019; Zhang et al. 2016]. Warping-based methods [Averbuch-Elor et al. 2017; Geng et al. 2018; Siarohin et al. 2019; Wiles et al. 2018; Yin et al. 2022] are typically few-shot, such that they require a source image or several frames as input. Give the source image, these methods do not reconstruct the face geometry but directly estimate the 2D warping from the source image to the target image. However, learning accurate facial warping across different head poses is very challenging. To alleviate this issue, face templates are also introduced and served as the guidance of the warping [Doukas et al. 2021; Ren et al. 2021]. Recently, 3D flow fields or volumetric features are also leveraged in [Drobyshev et al. 2022; Wang et al. 2021b] to generate avatar images under different head poses. On the contrary, mapping-based methods [Moser et al.

2021; Naruniec et al. 2020; Yan et al. 2018] directly learn the mapping between the face images of different subjects without using templates or warping maps. Despite the promising results achieved by the above image-based face reenactment methods, they require large-scale datasets for training and typically suffer from the lack of view consistency.

3 LATENT HEAD NERF

LatentAvatar consists of a latent head NeRF, which can be driven by a latent expression code. The latent expression code is learned in an end-to-end and self-supervised manner along with the head NeRF, for the goal of capturing person-specific detailed expressions.

3.1 Formulation

Previous work [Gafni et al. 2021; Gao et al. 2022; Guo et al. 2021; Liu et al. 2022] has exploited the power of NeRF [Mildenhall et al. 2020] for 3D head modeling. A typical head NeRF model Φ can be formulated as an expression-conditioned implicit field:

$$(c, \sigma) = \Phi(x, d, \theta), \quad (1)$$

where x and d are the query point and view direction used in volumetric rendering, c and σ are the color and the density respectively, and θ denotes the expression condition. To control the NeRF-based head model, previous methods [Gafni et al. 2021; Gao et al. 2022] typically use the expression coefficients of face templates such as 3DMM [Gerig et al. 2017] as the expression condition θ . In these methods, the expression coefficients are obtained by tracking face templates from the input images and are not learnable during the NeRF optimization. This process introduces two intractable limitations: i) the insufficient expression ability of 3DMM model itself makes it difficult to cover high-frequency details of person-specific expressions, and ii) the inaccuracy of the face tracking methods incurs additional deviations in the expression condition.

To tackle these issues, we propose a latent head NeRF, which does not use any pretrained face templates to model the expressions of head avatars but learns the expression condition in a latent space jointly with the NeRF optimization. Previous work [Hong et al. 2022] has tried naively learning the expression latent space in NeRF without constraints, but inversion is needed to retrieve the latent code needs and introduces additional difficulty in the avatar reenactment. In contrast, our method learns the latent code from face images and treats the latent head NeRF as an autoencoder. Specifically, the face image is firstly encoded into the latent expression space and then decoded via volume rendering of the radiance field. Formally, the latent head NeRF can be formulated as:

$$\theta_p = E(I_{face}), \text{ and} \quad (2)$$

$$(c, \sigma) = \Phi(x, d, \theta_p), \quad (3)$$

where E denotes the encoder, I_{face} denotes the face region of the portrait image, and θ_p denotes the person-specific latent expression code. Once the autoencoder is trained, we can simply feed face images of the avatar into the encoder and use the resulting latent expression code to drive the head NeRF avatar.

Compared with the previous head NeRF [Gafni et al. 2021; Gao et al. 2022], the expression condition of our head NeRF is a learnable latent code instead of the pre-defined expression coefficients of face templates. With the photometric reconstruction objectives, the

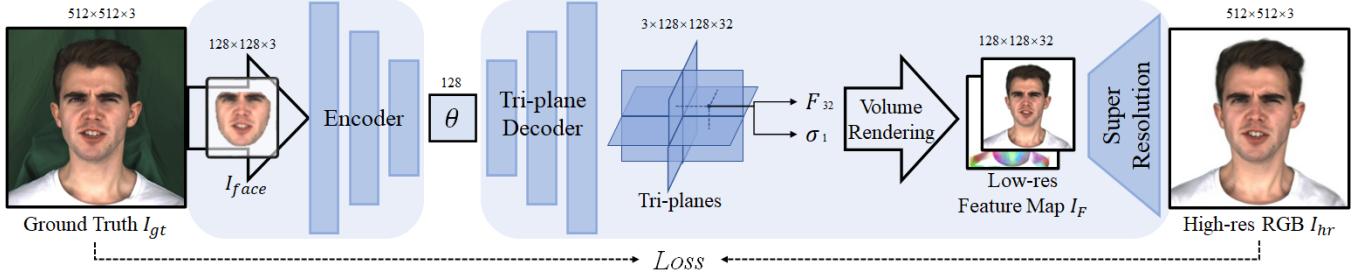


Figure 2: Overview of the Latent Head NeRF. Given a portrait video, we first encode the face image to the latent expression code θ , which is used as a condition to generate the tri-plane features. Given a 3D position, the feature vector H is extracted from the tri-plane features for the volume rendering of the low-resolution image and feature map. finally, a super-resolution network is used to generate the corresponding high-resolution images.

latent expression code is learned to capture the finer-grained details of person-specific expressions during the NeRF optimization. In our experiments, we will show that our latent expression code can even capture the subtle movement of teeth, tongue, and eyeballs of the subject, which goes far beyond the PCA-based expression coefficients of the face templates such as 3DMM [Gerig et al. 2017].

3.2 NeRF-based Decoder

In this part, we present the detailed architecture of our NeRF-based decoder Φ . In the generic dynamic NeRF [Gafni et al. 2021], the rendered images tend to be blurred due to the insufficient ability of pixel-wise feature learning. To tackle this, we explore a tri-plane representation and hybrid rendering [Chan et al. 2022] in our latent head NeRF. Specifically, we feed the latent expression code θ_p to a StyleGAN-based 2D convolutional network [Karras et al. 2021] to generate tri-plane features ($\mathcal{H}_{xy}, \mathcal{H}_{yz}, \mathcal{H}_{xz}$). Given a 3D position x , three feature vectors (H_{xy}, H_{yz}, H_{xz}) are queried by projecting it onto each of the three feature planes. Then, an aggregated feature vector H is obtained by summing up the three vectors. The feature vector H is further processed by a lightweight MLP for the generation of the color c , the density σ , and a high dimensional color feature F . Given a camera pose, we first render a low-resolution face image I_{lr} , a feature map I_F through the volume rendering based on the color c and feature F , respectively. Meanwhile, a mask map M with the same resolution is also generated during the rendering. Then, the low-resolution face image I_{lr} is first concatenated with the feature map I_F and fed into a super-resolution module to generate a high-resolution RGB portrait image I_{hr} . In contrast to the purely generative tasks in EG3D [Chan et al. 2022], our head avatar pays more attention to the rich expression enhancement of a specific person. Thus, the model size of the tri-plane generator can be largely reduced. In our solution, a U-net structure with downsampling layers is used in the super-resolution module.

3.3 Latent Code Optimization

For end-to-end learning of the latent expression code, we jointly optimize the parameters of the encoder and the decoder of the latent head NeRF. The total loss is :

$$\begin{aligned} \mathcal{L} = & ||I_{hr} - I_{gt}||_1 + \lambda_{vgg} VGG(I_{hr}, I_{gt}) \\ & + \lambda_{lr} ||I_{lr} - I_{gt}||_1 + \lambda_{mask} ||M - M_{gt}||_2, \end{aligned} \quad (4)$$

where I_{gt} and M_{gt} denote the preprocessed ground-truth image and mask respectively, I_{lr} denotes the first three channels of the low-resolution feature map I_F , M denotes the rendered low-resolution mask, I_{hr} denotes the final high-resolution image, and λ denotes the weight of each term. Besides, a VGG perceptual loss [Zhang et al. 2018] $VGG(\cdot)$ is also used during the training.

4 CROSS-IDENTITY REENACTMENT

Existing NeRF-based head avatar methods [Gafni et al. 2021; Gao et al. 2022; Xu et al. 2023] use the expression coefficients of templates as the drive signal, which can be easily used to achieve cross-identity reenactment. In our LatentAvatar, the drive signal is the latent expression code learned in a person-specific setting. The person-specific latent expression code of different subjects can not be exchanged for reenactment because they are learned separately. To enable cross-identity reenactment, we further leverage a Y-shaped network to learn a shared latent expression code and then map it to the person-specific one as the drive signal.

4.1 Y-shaped Network

Given monocular videos of the avatar and the actor, the shared latent expression code should allow face swapping of these two subjects while ensuring expression consistency. Inspired by previous work [Naruniec et al. 2020; Yan et al. 2018], we construct a network with a Y-shaped architecture, which contains a shared encoder E_{shared} and two separate decoders D_{ava}, D_{act} . As shown in Fig. 3, the shared encoder E_{shared} encodes the face images of both subjects into the **shared expression latent space** γ . Then, the avatar and actor decoders D_{ava}, D_{act} map the latent code to the face images of the two subjects individually.

The Y-shaped network is learned with the reconstruction task of both the actor and avatar images. For more efficient learning, we only optimize the parameters of the shared encoder E_{shared} and the selected decoder of the avatar or actor, while freezing the parameters of another decoder. More specifically, in each iteration during the training phase, we randomly select one subject for training. Here, let the selected subject be the actor for a simplified explanation. The face region image I_{act} is fed to the shared encoder E_{shared} to produce the latent expression code γ_{act} in the shared latent space. Then, the latent code γ_{act} is fed into the decoder D_{act}

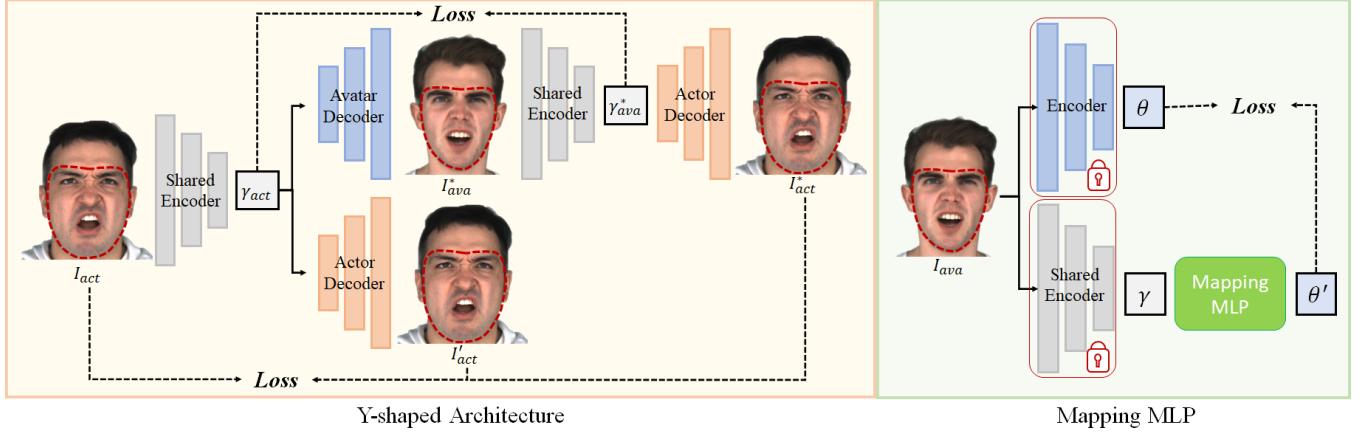


Figure 3: Illustration of the Y-shape network (left) used to learn the shared latent expression code and the mapping MLP (right) used to map the shared latent code to its person-specific one. In the Y-shaped network, the shared encoder E_{shared} encodes the input face images as the shared expression latent code, which will be decoded by the avatar and actor decoders D_{ava}, D_{act} to the face images of the avatar and the actor individually. To bridge the shared and person-specific latent space, the mapping MLP learns to map the shared latent expression code to the person-specific one.

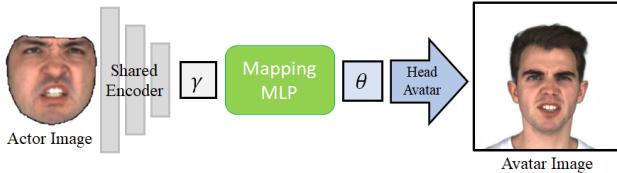


Figure 4: The process of the cross-identity reenactment in our method. The face image of the actor is first fed into the shared encoder to obtain the shared latent code γ , which is mapped as the person-specific latent code θ to drive the NeRF-based head avatar.

for the generation of the face image I'_{act} , which will be optimized by the self-reconstruction loss function:

$$\mathcal{L}_{rec} = ||I'_{act} - I_{act}||_1 + \lambda_{ssim} SSIM(I'_{act}, I_{act}), \quad (5)$$

where $SSIM(\cdot)$ denotes Structure Similarity Index and λ_{ssim} denotes the weight.

In the original Y-shaped network [Yan et al. 2018], there is no restriction imposed on the learning of the shared latent space, which may lead to a separate distribution of the latent codes of input subjects. However, to ensure the expression consistency of different subjects, the distributions of the shared latent code should overlap with each other as much as possible for the two input subjects. Inspired by CycleGAN [Zhu et al. 2017], we leverage a similar cycle consistency loss to guarantee this. Specifically, the actor latent code γ_{act} is fed into the avatar decoder D_{ava} to generate the avatar image I_{ava}^* , which will be further fed into the shared encoder E_{shared} to obtain a new avatar latent code γ_{ava}^* . By taking γ_{ava}^* as input, the actor decoder can generate another actor image I_{act}^* . The cycle consistency loss is imposed on the newly generated latent code and

images, which can be formulated as:

$$\begin{aligned} \mathcal{L}_{cycle} = & ||I_{act}^* - I_{act}||_1 + \lambda_{ssim} SSIM(I_{act}^*, I_{act}) \\ & + \lambda_{code} ||\gamma_{ava}^* - \gamma_{act}||_2. \end{aligned} \quad (6)$$

Overall, the total loss function can be formulated as:

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{cycle}. \quad (7)$$

4.2 Mapping MLP

To enable the cross-identity reenactment of LatentAvatar, we additionally train a small mapping MLP to bridge the shared and person-specific latent codes θ and γ . Specifically, this mapping MLP maps the latent expression code in the shared latent space to the one in the person-specific latent space. In this way, the input data just needs to go through only one encoder E_{shared} and a small mapping MLP to obtain the drive signal (i.e., the latent expression code θ) for cross-identity reenactment, as shown in Fig. 4.

To train the mapping MLP, we only need the training data of the avatar. During training, we freeze all the parameters in the previously trained autoencoder module and the Y-shaped network and only optimize the parameters of the mapping MLP. Given a face image of the avatar I_{ava} , we encode it to two latent codes: the shared latent expression code γ and the person-specific latent expression code θ through the corresponding encoders E_{shared} and E , respectively. Then, the mapping MLP is used to map the shared latent code to the person-specific one. The loss function can be formulated as:

$$\mathcal{L} = ||E(I_{ava}) - Mapping(E_{shared}(I_{ava}))||_2, \quad (8)$$

where $Mapping(\cdot)$ denotes the mapping MLP.

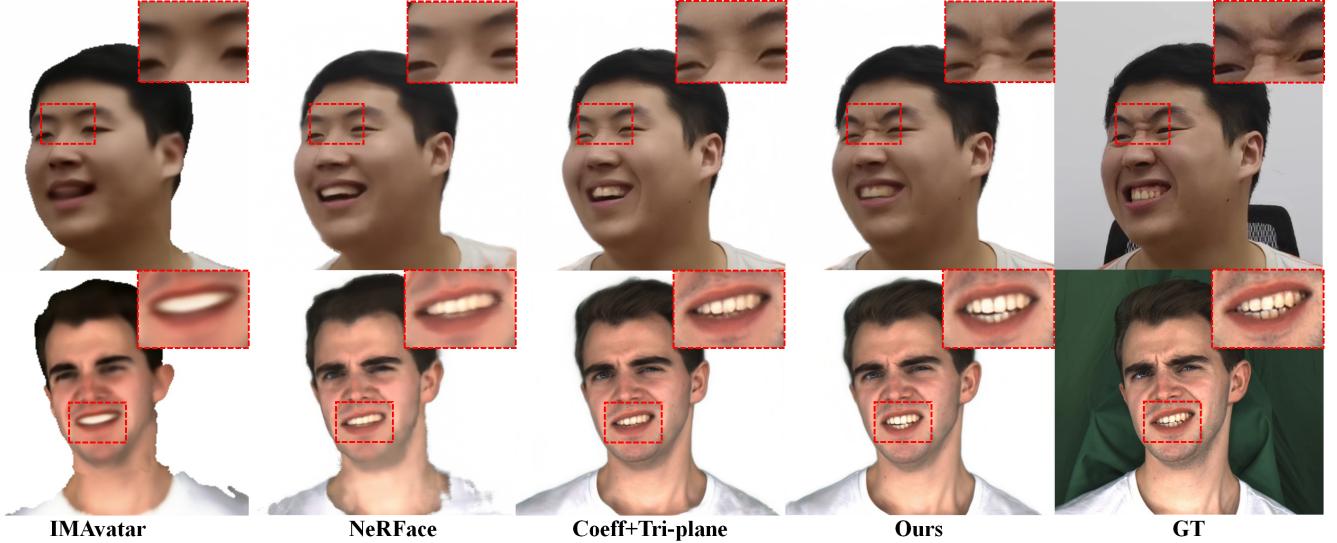


Figure 5: Qualitative comparisons of different methods on the self reenactment task. From left to right: IMAvatar [Zheng et al. 2022], NeRFace [Gafni et al. 2021], Coeff+Tri-plane, and Ours. Our method surpasses other methods in the ability to capture and reproduce detailed expressions such as the wrinkles around the nose and the exposure level of teeth.

5 EXPERIMENTS

5.1 Implementation Details

The training of our LatentAvatar requires the video data of different subjects. Specifically, we use the videos from a public dataset MEAD [Wang et al. 2020] and our portrait video dataset collected by a hand-held mobile phone. For MEAD, we selected 3 representative subjects. Each video contains about 60,000 frames. For our dataset, we capture videos of 4 subjects. Each video contains about 10,000 frames. We trim 10% fragment from the begining or the end of each video for evaluation and the remaining 90% for training. During the data preprocessing phase, all videos are resized to 512×512 resolution. We also follow previous work [Ke et al. 2020; Lin et al. 2022] to remove the background from each video frame to obtain ground truth portrait images for training.

In our experiments, all input images of the encoder and the shared encoder are resized to the resolution of 128×128 . Moreover, following the pre-processing strategy in DeepFaceLab [Perov et al. 2021], we localize the face region using the detected 68 face landmarks so that the input images of the encoder mainly contain the valid face region. The dimensions of the person-specific and shared latent expression code are set as 128 and 256, respectively. The resolutions of both the tri-plane features and the low-resolution feature map are set as 128×128 while their channel number is set as 32.

During optimization, we use an Adam [Kingma and Ba 2017] optimizer. The learning rate is set as 1×10^{-4} for all learnable parameters. For ray sampling in volumetric rendering, we sample 64 points along each ray in each iteration. The latent head NeRF model is trained for 200,000 iterations with a batch size of 2, while the Y-shaped architecture and the mapping MLP are trained for 30,000 iterations with a batch size of 16. The weights of different

loss terms are set as follows: $\lambda_{vgg} = 0.1$, $\lambda_{lr} = 0.1$, $\lambda_{mask} = 0.1$, $\lambda_{ssim} = 0.1$ and $\lambda_{code} = 1 \times 10^{-4}$.

5.2 Results and Comparisons

The key idea of our LatentAvatar is to use the latent expression code as the drive signal to animate the head avatar. To validate its efficacy, we build an ablation baseline method named Coeff+Triplane. In Coeff+Triplane, the architecture of the head NeRF module is exactly the same as our method, i.e., using the tri-plane representation and the super-resolution module to improve the quality of the generated images. Just as NeRFace [Gafni et al. 2021], Coeff+Triplane directly uses the 3DMM expression coefficients as the drive signal, which is the only difference in comparison with our method.

5.2.1 Self Reenactment. We conduct qualitative and quantitative comparisons between our method, Coeff+Tri-plane, and other two state-of-the-art methods, i.e., IMAvatar [Zheng et al. 2022], and NeRFace [Gafni et al. 2021], on our dataset. IMAvatar reconstructs an implicit represented [Yariv et al. 2020] head avatar based on FLAME model [Li et al. 2017], while NeRFace reconstructs a dynamic NeRF represented head avatar with 3DMM expression coefficients as the drive signal.

The self reenactment results of different methods are shown in Fig. 5. We can see that the accuracy of the template fitting is crucial for IMAvatar due to the reliance on face templates in its reconstruction process. The performance of IMAvatar is inferior to our method, especially on our newly collected video data. Our video data contains a large number of challenging and dramatic expressions, resulting in inaccurate landmark detection results and the failure of template fitting. As a result, the texture of the reconstructed avatar of IMAvatar tends to be much more blurred. Though NeRFace and Coeff+Tri-plane do not use the geometry or texture

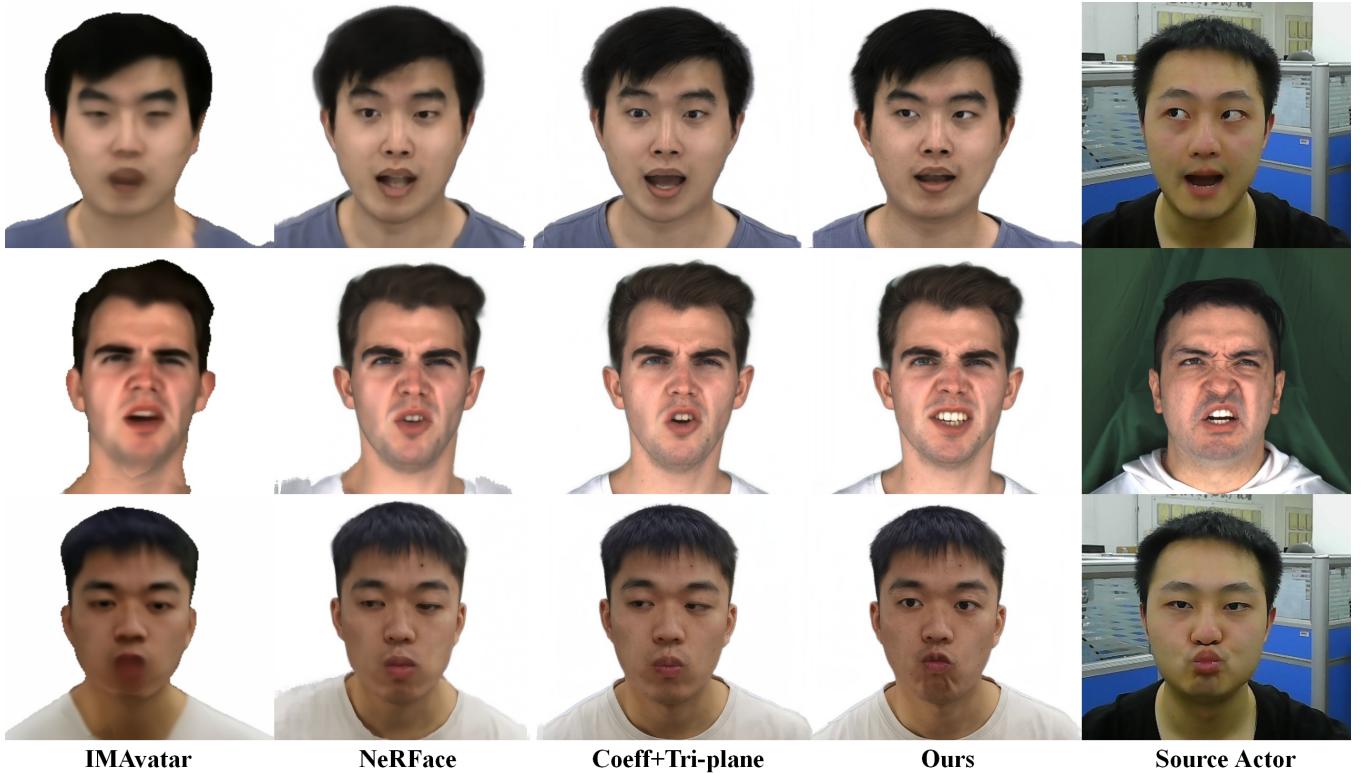


Figure 6: Qualitative comparisons of different methods on the cross-identity reenactment task. From left to right: IMAvatar, NeRFace, Coeff+Tri-plane baseline and Ours. Our method can accurately transfer eye movement and tooth grinning and remain robust in some exaggerated expressions.

of the face template, they still rely on the 3DMM expression coefficients. As shown in Fig. 5, their reconstructed head avatar is not able to produce detailed wrinkles (see the first row) or eye movements (see the second row) since these subtle expressions cannot be explicitly represented by the face template. In contrast, the drive signal of our method is the latent expression code extracted from the input face image directly. Such a learnable expression code is able to capture high-frequency and detailed expression information, which goes far beyond the expression ability of existing face templates.

For quantitative evaluation, all the methods are quantitatively evaluated by calculating Mean Square Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structure Similarity Index (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) [Zhang et al. 2018] between the ground truth image and the generated image on the evaluation data. During the evaluation, we select 6 avatars to perform self reenactment and calculate the mean values of the evaluation results. The numerical results are reported in Tab. 1 for comparisons. We can see that our method convincingly outperforms all other state-of-the-art methods.

5.2.2 Cross-identity Reenactment. We show the cross-identity reenactment results of different methods in Fig. 6. For NeRFace and Coeff+Tri-plane, we can observe that these two methods may produce unreasonable results during the expression transfer between

Table 1: Quantitative evaluation results of our method, Coeff+Tri-plane (baseline), IMAvatar [Zheng et al. 2022], and NeRFace [Gafni et al. 2021].

Method	MSE $\times 10^{-3}$ ↓	PSNR ↑	SSIM ↑	LPIPS ↓
IMAvatar	6.89	21.79	0.871	0.209
NeRFace	3.39	25.64	0.903	0.135
Coeff+Tri-plane	2.70	27.00	0.917	0.049
Ours	2.61	27.61	0.919	0.048

different subjects. These can be explained by the fact that the identity coefficients and the expression coefficients of the face templates are easy to be coupled with each other. In addition, these template-based methods cannot produce avatar images with challenging expressions such as pouting when the corresponding landmark shape is difficult to be detected. As shown in Fig. 6, our method performs well under challenging cases and even captures the details with tooth baring (see the 2nd row), while the template-based methods suffer from serious artifacts when the expression coefficients are out of the distributions in the training dataset (see the 3rd row). The success of our method can be attributed to the expression codes learned in the person-specific and shared latent space. Both of these two latent codes and their mapping are learned in an end-to-end and self-supervised manner, enabling the capture

of subtle expressions and their correspondences between different subjects. Moreover, our method does not rely on template fitting and hence enjoys both of expressiveness and stability.

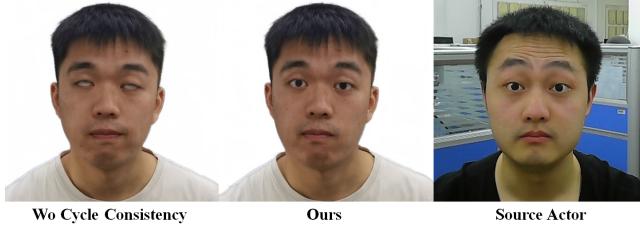


Figure 7: Ablation study on cycle consistency loss

5.2.3 Ablation Study on Cycle Consistency Loss. We compared our training pipeline with and without cycle consistency loss in cross-identity reenactment experiments. Qualitative results are shown in Fig. 7. Since there is no cycle consistency loss to constraint, the latent codes are more diffuse in the shared latent space, resulting in the expression consistency reducing significantly.

6 CONCLUSION

In this work, we have proposed LatentAvatar, an expressive NeRF-based head avatar driven by latent expression codes. LatentAvatar leverages a latent head NeRF and a Y-shaped network to learn the latent expression code in the person-specific and shared space respectively. These two types of latent expression code are further bridged by a mapping MLP to achieve cross-identity reenactment. Experimental results have demonstrated the capability of our method to capture detailed expressions and subtle movements of the teeth and eyeballs, which shows significant improvement over previous solutions. Moreover, LatentAvatar gets rid of the template tracking issues and hence is more robust and stable to challenging expressions. We believe that the combination of NeRF representations and the learned latent expression code is a promising direction to achieve lightweight and expressive head avatar reconstruction and reenactment.

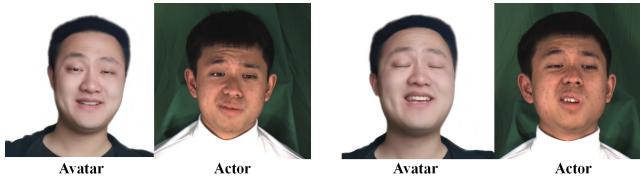


Figure 8: Failure cases when there are distinct differences between the appearance and expression distribution of the two identities.

7 DISCUSSION

Ethical Considerations. Our method can synthesize photo-realistic fake portrait videos, but also brings a serious problem that it can be used to spread false information, manipulate public opinion, and undermine trust in media, leading to significant societal harm. Therefore, an efficient and accurate method for discriminating forgery is the most worthy of consideration.

Limitation. Despite the expressive results, there are two main limitations of our method. First, our method requires the learning of latent codes in person-specific and shared spaces, which implies that our method needs additional training on the video of the actor for the task of cross-identity reenactment. Second, when there are distinct differences between the appearance or expression distribution of the two identities, the shared latent code is inclined to contain individual appearance information, which may lead to the erroneous mapping of expressions. Two failure cases are shown in Fig. 8 for illustration.

Future Work. In future work, to tackle the above two issues, we may adopt meta-learning techniques and leverage existing large-scale video datasets to enhance the generalization of our solution to diverse facial appearances and expressions.

ACKNOWLEDGMENTS

This paper is supported by National Key R&D Program of China (2022YFF0902200), the NSFC project No.62125107 and No.61827805.

REFERENCES

- ShahRukh Athar, Zhixin Shu, and Dimitris Samaras. 2023. Flame-in-nerf: Neural control of radiance fields for free view face animation. In *IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. 1–8.
- ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. 2022. RigeRF: Fully Controllable Neural 3D Portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F. Cohen. 2017. Bringing Portraits to Life. *ACM Trans. Graph.* 36, 6, Article 196 (nov 2017), 13 pages.
- Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih. 2021. Deep Relightable Appearance Models for Animatable Faces. *ACM Trans. Graph.* 40, 4, Article 89 (jul 2021), 15 pages.
- Chen Cao, Vasu Agrawal, Fernando De La Torre, Lele Chen, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2021. Real-Time 3D Neural Facial Animation from Binocular Video. *ACM Trans. Graph.* 40, 4, Article 87 (jul 2021), 17 pages.
- Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. 2015. Real-Time High-Fidelity Facial Performance Capture. *ACM Trans. Graph.* 34, 4, Article 46 (jul 2015), 9 pages.
- Chen Cao, Qiming Hou, and Kun Zhou. 2014a. Displaced Dynamic Expression Regression for Real-Time Facial Tracking and Animation. *ACM Trans. Graph.* 33, 4, Article 43 (jul 2014), 10 pages.
- Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shouou-I Yu, Yaser Sheikh, and Jason Saragih. 2022. Authentic Volumetric Avatars from a Phone Scan. *ACM Trans. Graph.* 41, 4, Article 163 (jul 2022), 19 pages.
- Chen Cao, Yanlin Weng, Stephen Lin, and Kun Zhou. 2013. 3D Shape Regression for Real-Time Facial Animation. *ACM Trans. Graph.* 32, 4, Article 41 (jul 2013), 10 pages.
- Chen Cao, Yanlin Weng, Shun Zhou, Y. Tong, and Kun Zhou. 2014b. FaceWarehouse: A 3D Facial Expression Database for Visual Computing. In *IEEE Transactions on Visualization and Computer Graphics*, Vol. 20. 413–425.
- Chen Cao, Hongzhi Wu, Yanlin Weng, Tianjia Shao, and Kun Zhou. 2016. Real-Time Facial Animation with Image-Based Dynamic Avatars. *ACM Trans. Graph.* 35, 4, Article 126 (jul 2016), 12 pages.
- Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. 2022. Efficient Geometry-aware 3D Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16102–16112.
- Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. 2021. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14124–14133.
- Zhuo Chen, Chaoyue Wang, Bo Yuan, and Dacheng Tao. 2020. PuppeteerGAN: Arbitrary Portrait Animation With Semantic-Aware Appearance Transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13515–13524.
- Hang Chu, Shugao Ma, Fernando Torre, Sanja Fidler, and Yaser Sheikh. 2020. Expressive Telepresence via Modular Codec Avatars. In *Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV)*. 330–345.

- Kevin Dale, Kalyan Sunkavalli, Micah K. Johnson, Daniel Vlasic, Wojciech Matusik, and Hanspeter Pfister. 2011. Video Face Replacement. *ACM Trans. Graph.* 30, 6 (dec 2011), 1–10.
- Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. 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 (CVPR) Workshops*.
- Michail Christos Doukas, Mohammad Rami Koujan, Viktoria Sharmancka, Anastasios Roussos, and Stefanos Zafeiriou. 2020. Head2Head++: Deep Facial Attributes Re-Targeting. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3 (2020), 31–43.
- Michail Christos Doukas, Stefanos Zafeiriou, and Viktoria Sharmancka. 2021. HeadGAN: One-shot Neural Head Synthesis and Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Nikita Drobyshev, Jenya Chelishev, Taras Khakhlun, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. 2022. MegaPortraits: One-shot Megapixel Neural Head Avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*.
- Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. 2022. Fast Dynamic Radiance Fields with Time-Aware Neural Voxels. In *SIGGRAPH Asia 2022 Conference Papers*.
- Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Niessner. 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)*. 8645–8654.
- Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. 2022. Reconstructing Personalized Semantic Facial NeRF Models From Monocular Video. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* 41, 6 (2022).
- Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. 2018. Warp-Guided GANs for Single-Photo Facial Animation. *ACM Trans. Graph.* 37, 6, Article 231 (dec 2018), 12 pages.
- Thomas Gerig, Andreas Forster, Clemens Blumer, Bernhard Egger, Marcel Lüthi, Sandro Schönborn, and Thomas Vetter. 2017. Morphable Face Models - An Open Framework. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 75–82.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Conference on Neural Information Processing Systems (NeurIPS)*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (Eds.), Vol. 27. Curran Associates, Inc.
- Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. 2022. Neural Head Avatars From Monocular RGB Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18632–18643.
- Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. 2021. AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5764–5774.
- Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. 2022. HeadNeRF: A Real-Time NeRF-Based Parametric Head Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20374–20384.
- Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. 2017. Avatar Digitization from a Single Image for Real-Time Rendering. *ACM Trans. Graph.* 36, 6, Article 195 (nov 2017), 14 pages.
- Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. 2015. Dynamic 3D Avatar Creation from Hand-Held Video Input. *ACM Trans. Graph.* 34, 4, Article 45 (jul 2015), 14 pages.
- Tero Karras, Samuli Laine, and Timo Aila. 2021. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 12 (2021), 4217–4228.
- Zhanghan Ke, Jiayi Sun, Kaican Li, Qiong Yan, and Rynson W. H. Lau. 2020. MODNet: Real-Time Trimap-Free Portrait Matting via Objective Decomposition. In *AAAI Conference on Artificial Intelligence*.
- Taras Khakhlun, Vanessa Solyarova, Victor Lempitsky, and Egor Zakharov. 2022. Realistic One-shot Mesh-based Head Avatars. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Hyeyoungwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. 2018. Deep Video Portraits. *ACM Trans. Graph.* 37, 4, Article 163 (jul 2018), 14 pages.
- Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs.LG]*
- Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. 2017. Fast Face-Swap Using Convolutional Neural Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 3697–3705.
- Mohammad Rami Koujan, Michail Christos Doukas, Anastasios Roussos, and Stefanos Zafeiriou. 2020. Head2Head: Video-based Neural Head Synthesis. *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 16–23.
- Kai Li, Feng Xu, Jue Wang, Qionghai Dai, and Yebin Liu. 2012. A data-driven approach for facial expression synthesis in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 299–310.
- Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. 2017. Learning a Model of Facial Shape and Expression from 4D Scans. *ACM Trans. Graph.* 36, 6, Article 194 (nov 2017), 17 pages.
- Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. 2022. Robust High-Resolution Video Matting With Temporal Guidance. In *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 3132–3141.
- Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. 2022. Semantic-Aware Implicit Neural Audio-Driven Video Portrait Generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep Appearance Models for Face Rendering. *ACM Trans. Graph.* 37, 4, Article 68 (July 2018), 13 pages.
- Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural Volumes: Learning Dynamic Renderable Volumes from Images. *ACM Trans. Graph.* 38, 4, Article 65 (July 2019), 14 pages.
- Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. 2021. Mixture of Volumetric Primitives for Efficient Neural Rendering. *ACM Trans. Graph.* 40, 4, Article 59 (jul 2021), 13 pages.
- Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. 2021. Pixel Codec Avatars. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 64–73.
- Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy Networks: Learning 3D Reconstruction in Function Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Lucio Moser, Chinny Chien, Mark Williams, Jose Serra, Darren Hendler, and Doug Roble. 2021. Semi-Supervised Video-Driven Facial Animation Transfer for Production. *ACM Trans. Graph.* 40, 6, Article 222 (dec 2021), 18 pages.
- Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multi-resolution Hash Encoding. *ACM Trans. Graph.* 41, 4, Article 102 (July 2022), 15 pages.
- Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. 2018. PaGAN: Real-Time Avatars Using Dynamic Textures. *ACM Trans. Graph.* 37, 6, Article 258 (dec 2018), 12 pages.
- Jacek Naruniec, Leonhard Helminger, Christopher Schroers, and Romann M. Weber. 2020. High-Resolution Neural Face Swapping for Visual Effects. *Computer Graphics Forum* 39, 4 (2020), 173–184.
- Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. 2018. FSNet: An Identity-Aware Generative Model for Image-based Face Swapping. In *Asian Conference on Computer Vision*.
- Yuval Nirkin, Yosi Keller, and Tal Hassner. 2019a. FSGAN: Subject Agnostic Face Swapping and Reenactment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 7183–7192.
- Yuval Nirkin, Yosi Keller, and Tal Hassner. 2019b. FSGAN: Subject Agnostic Face Swapping and Reenactment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 7183–7192.
- Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. 2018. On Face Segmentation, Face Swapping, and Face Perception. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. 98–105.
- Kyle Olszewski, Zimo Li, Chao Yang, Yi Zhou, Ronald Yu, Zeng Huang, Sitao Xiang, Shunsuke Saito, Pushmeet Kohli, and Hao Li. 2017. Realistic Dynamic Facial Textures From a Single Image Using GANs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5439–5448.
- Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 165–174.
- Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. 2021a. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5845–5854.
- Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. 2021b. HyperNeRF: A Higher-Dimensional Representation for Topologically Varying Neural Radiance Fields. *ACM Trans. Graph.* 40, 6, Article 238 (dec 2021).
- Ivan Perov, Daiheng Gao, Nikolay Chervonyi, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr. Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, Pingyu Wu, Bo Zhou, and Weiming Zhang. 2021. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. *arXiv:2005.05535 [cs.CV]*

- Yurui Ren, Ge Li, Yuanqi Chen, Thomas H. Li, and Shan Liu. 2021. PIRenderer: Controllable Portrait Image Generation via Semantic Neural Rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 13759–13768.
- Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Conference on Neural Information Processing Systems (NeurIPS)*.
- Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. 2022a. IDE-3D: Interactive Disentangled Editing for High-Resolution 3D-aware Portrait Synthesis. *ACM Transactions on Graphics (TOG)* 41, 6, Article 270 (2022), 10 pages.
- Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. 2023. Next3D: Generative Neural Texture Rasterization for 3D-Aware Head Avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. 2022b. FENeRF: Face Editing in Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7662–7672.
- Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. 2015. Real-Time Expression Transfer for Facial Reenactment. *ACM Trans. Graph.* 34, 6, Article 183 (oct 2015), 14 pages.
- Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Niessner. 2016. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2387–2395.
- Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. 2005. Face Transfer with Multilinear Models. *ACM Trans. Graph.* 24, 3 (jul 2005), 426–433.
- Daoyu Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo Gotardo. 2022. MoRF: Morphable Radiance Fields for Multiview Neural Head Modeling. In *ACM SIGGRAPH 2022 Conference Proceedings* (Vancouver, BC, Canada) (*SIGGRAPH '22*). Association for Computing Machinery, New York, NY, USA, Article 55, 9 pages.
- Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020. MEAD: A Large-scale Audio-visual Dataset for Emotional Talking-face Generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Lizhen Wang, Xiaochen Zhao, Jingxiang Sun, Yuxiang Zhang, Hongwen Zhang, Tao Yu, and Yebin Liu. 2023. StyleAvatar: Real-time Photo-realistic Portrait Avatar from a Single Video. In *ACM SIGGRAPH 2023 Conference Proceedings*.
- Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021b. One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10034–10044.
- Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhofer. 2021a. Learning Compositional Radiance Fields of Dynamic Human Heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5704–5713.
- Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. 2011. Realtime Performance-Based Facial Animation. *ACM Trans. Graph.* 30, 4, Article 77 (jul 2011), 10 pages.
- Olivia Wiles, A. Sophia Koepke, and Andrew Zisserman. 2018. X2Face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Yuelang Xu, Lizhen Wang, Xiaochen Zhao, Hongwen Zhang, and Yebin Liu. 2023. AvatarMAV: Fast 3D Head Avatar Reconstruction Using Motion-Aware Neural Voxels. In *ACM SIGGRAPH 2023 Conference Proceedings*.
- Shuqi Yan, Shaorong He, Xue Lei, Guanhua Ye, and Zhipeng Xie. 2018. Video Face Swap Based on Autoencoder Generation Network. *2018 International Conference on Audio, Language and Image Processing (ICALIP)*, 103–108.
- Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. 2020. Multiview Neural Surface Reconstruction by Disentangling Geometry and Appearance. In *Conference on Neural Information Processing Systems (NeurIPS)*, Vol. 33.
- T Yenamandra, A Tewari, F Bernard, HP Seidel, M Elgarhib, D Cremers, and C Theobalt. 2021. i3DMM: Deep Implicit 3D Morphable Model of Human Heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Fei Yin, Yong Zhang, Xiaodong Cun, Ming Cao, Yanbo Fan, Xuanxia Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. 2022. StyleHEAT: One-Shot High-Resolution Editable Talking Face Generation via Pre-trained StyleGAN. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelNeRF: Neural Radiance Fields from One or Few Images. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4576–4585.
- Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. 2019. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 9458–9467.
- Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. 2022. FDNeRF: Few-Shot Dynamic Neural Radiance Fields for Face Reconstruction and Expression Editing. In *SIGGRAPH Asia 2022 Conference Papers* (Daegu, Republic of Korea) (*SA '22*). Association for Computing Machinery, New York, NY, USA, Article 12, 9 pages.
- Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.
- Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 586–595.
- Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. 2022. I M Avatar: Implicit Morphable Head Avatars From Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13535–13545.
- Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. 2023a. PointAvatar: Deformable Point-based Head Avatars from Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. 2023b. AvatarReX: Real-time Expressive Full-body Avatars. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–19. <https://doi.org/10.1145/3592101>
- Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2242–2251.
- Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. 2022. MoFaNeRF: Morphable Facial Neural Radiance Field. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Wojciech Zienonka, Timo Bolkart, and Justus Thies. 2022. Instant Volumetric Head Avatars. arXiv:2211.12499 [cs.CV]