# Video-driven Neural Physically-based Facial Asset for Production


LONGWEN ZHANG[*], ShanghaiTech University and Deemos Technology Co., Ltd., China
CHUXIAO ZENG[*], ShanghaiTech University and Deemos Technology Co., Ltd., China
QIXUAN ZHANG[*], ShanghaiTech University and Deemos Technology Co., Ltd., China
HONGYANG LIN, ShanghaiTech University and Deemos Technology Co., Ltd., China
RUIXIANG CAO, ShanghaiTech University and Deemos Technology Co., Ltd., China
WEI YANG, Huazhong University of Science and Technology, China
LAN XU[†], ShanghaiTech University and Shanghai Engineering Research Center of Intelligent Vision and Imaging, China
JINGYI YU[†], ShanghaiTech University and Shanghai Engineering Research Center of Intelligent Vision and Imaging, China


Fig. 1. Top: Our neural physically-based facial assets with stylized appearance. Bottom: cross-identity facial motion re-targeting driven by phone videos.


Production-level workflows for producing convincing 3D dynamic human faces have long relied on an assortment of labor-intensive tools for geometry and texture generation, motion capture and rigging, and expression synthesis. Recent neural approaches automate individual components but the corresponding latent representations cannot provide artists with explicit controls as in conventional tools. In this paper, we present a new learning-based, video-driven approach for generating dynamic facial geometries with high-quality physically-based assets. For data collection, we construct a hybrid multiview-photometric capture stage, coupling with ultra-fast video cameras to obtain raw 3D facial assets. We then set out to model the facial expression, geometry and physically-based textures using separate VAEs where we impose a global MLP based expression mapping across the latent spaces of respective networks, to preserve characteristics across respective attributes. We also model the delta information as wrinkle maps for the physically-based textures, achieving high-quality 4K dynamic textures. We demonstrate our approach in high-fidelity performer-specific facial capture and cross-identity facial motion retargeting. In addition, our multi-VAE-based neural asset, along with the fast adaptation schemes, can also be deployed to handle in-the-wild videos. Besides, we motivate the utility of our explicit facial disentangling strategy by providing various promising physically-based editing results with high realism. Comprehensive experiments show that our technique provides higher accuracy and visual fidelity than previous video-driven facial reconstruction and animation methods.



CCS Concepts: • **Computing methodologies** → **Motion capture**.

Additional Key Words and Phrases: Physically-Based Face Rendering, Facial Modeling, Digital Human, Video-Driven Animation



[*]Equal contributions.
[†]Corresponding author.

Authors' addresses: Longwen Zhang, ShanghaiTech University and Deemos Technology Co., Ltd., Shanghai, China, zhanglw2@shanghaitech.edu.cn; Chuxiao Zeng, ShanghaiTech University and Deemos Technology Co., Ltd., Shanghai, China, zengchx@shanghaitech.edu.cn; Qixuan Zhang, ShanghaiTech University and Deemos Technology Co., Ltd., Shanghai, China, zhangqx1@shanghaitech.edu.cn; Hongyang Lin, ShanghaiTech University and Deemos Technology Co., Ltd., Shanghai, China, linhy@shanghaitech.edu.cn; Ruixiang Cao, ShanghaiTech University and Deemos Technology Co., Ltd., Shanghai, China, caorx@shanghaitech.edu.cn; Wei Yang, Huazhong University of Science and Technology, Wuhan, China, weiyangcs@hust.edu.cn; Lan Xu, ShanghaiTech University and Shanghai Engineering Research Center of Intelligent Vision and Imaging, Shanghai, China, xulan1@shanghaitech.edu.cn; Jingyi Yu, ShanghaiTech University and Shanghai Engineering Research Center of Intelligent Vision and Imaging, Shanghai, China, yujingyi@shanghaitech.edu.cn.




0730-0301/2022/12-ART208
https://doi.org/10.1145/3550454.3555445

**ACM Reference Format:**
Longwen Zhang, Chuxiao Zeng, Qixuan Zhang, Hongyang Lin, Ruixiang Cao, Wei Yang, Lan Xu, and Jingyi Yu. 2022. Video-driven Neural Physically-based Facial Asset for Production . *ACM Trans. Graph.* 41, 6, Article 208 (December 2022), 16 pages. https://doi.org/10.1145/3550454.3555445

## 1 INTRODUCTION

High-quality and realistic digital avatars have been increasingly deployed in feature films, game productions, the entertainment industry, and most recently immersive experiences in the Metaverse. Despite tremendous advances in modeling 3D human body shapes [Loper et al. 2015], generating 3D dynamic facial assets at a production quality remains challenging: we humans are extremely sensitive to the perception of facial idiosyncrasies where even the slightest inconsistency can be easily detected. Successful solutions require not only reconstructing high-fidelity geometry, texture maps, normal maps, etc as in static facial asset generation but also reproducing realistic facial movement simulations indistinguishable from real videos.

To avoid the uncanny valley, production-level workflows commonly employ an assortment of tools for building compelling and convincing dynamic facial assets. They range from scanning real 3D faces via expensive apparatus such as a Light Stage [Debevec et al. 2000] or a multi-camera dome [Joo et al. 2017], to conducting physically-based simulations to faithfully reproduce how light interacts with the soft tissues of the face in terms of appearance attributes (e.g., diffuse albedo, specular intensity, roughness, normal/displacement maps) [Li et al. 2020a], and to capturing fine-grained facial motions [Laine et al. 2017] and transferring them onto a new performer with different physiognomy while performing a wide range of expressions [Moser et al. 2021]. Once collected, the rich raw data at individual stages need to undergo time-consuming manual post-processing steps, where highly skilled artists need to carefully adjust both the physically-based textures and the facial expression models, e.g., following the Facial Action Coding System (FACS) [Ekman and Friesen 1978]. The standard pipeline is thus labor-intensive and expensive (in both cost and time) to be deployed to support a broader audience. To facilitate easier dynamic facial asset generations, there is an urgent demand to develop a more automatic workflow to at least reduce, if not fully eliminate, the manual labor in the production cycle.

Recent deep learning techniques [Lattas et al. 2021; Li et al. 2020b,a] have provided a viable path to automated facial asset generation directly from the captured data with minimal human intervention. In particular, the data-driven strategy employed by learning-based techniques is applicable to handling datasets of different scales, as many as thousands of subjects for producing a generic model or as few as a single subject with different expressions for portrait animation. For the former, generic models [Bao et al. 2021; Feng et al. 2021; Li et al. 2020b, 2017; Raj et al. 2021] achieve high robustness despite environment lighting changes and variations in facial geometry and motion. Their generated assets are readily deployed to the existing facial animation production pipeline but they still rely on accurate blendshapes to achieve explicit control, which often lead to loss of fine-grained details and cannot meet production-level quality. For the latter, recent performer-specific neural modeling tools can produce both geometry and appearance with fine details and conduct compelling end-to-end facial rendering [Cao et al. 2021; Gafni et al. 2021; Lombardi et al. 2018, 2019, 2021; Wei et al. 2019] at photo-realism. However, unlike generic models, neural approaches as an implicit representation cannot yet provide explicit controls over fine details in geometry, motion or physically-based textures, and therefore cannot be readily integrated into existing production pipelines for games or feature films, which undermines the efficacy and efficiency of producing realistic and controllable facial assets.

In this paper, we present a novel learning-based, video-driven approach for generating dynamic facial geometry along with high-quality physically-based textures including pore-level diffuse albedo, specular intensity and normal maps. Our approach supports both high-fidelity performer-specific facial capture and cross-identity facial motion retargeting. It is directly applicable for producing various physically-based effects such as geometry and material editing, facial feature (e.g., wrinkle) transfer, etc, as shown in Fig. 1.

We recognize that using videos, instead of separate shots, to produce physically-based facial assets imposes several challenges. On the data front, by far there are very few high-quality 3D facial video datasets publicly available. On processing, there lacks tailored neural modeling algorithms to handle the spatial-temporal facial data. On editing, companion tools need to be developed. We hence first construct a multi-view photometric fast capture stage or FaStage that extends previous solutions such as the USC Light Stage [Debevec et al. 2000] to acquire facial geometry and textures under fast motion. FaStage employs a high-speed camera dome with 24 2K video cameras, a dense set of space-and-time-multiplexed illumination array, and three ultra-fast video cameras (Flex4K). The dome system is synchronized with the illumination pattern for obtaining proxy geometry whereas the Flex4K camera triplet can capture 38 gradient illumination patterns for normal and texture recovery at an ultra-high quality.

We then apply Variational Autoencoders (VAE) [Kingma and Welling 2014] to model the facial expression, geometry and physically-based textures of our assets via a unified neural representation. Different from prior work [Lombardi et al. 2018; Moser et al. 2021], we introduce a multi-VAE framework to model the characteristics of individual facial attributes and maintain a well-structured latent space, so that the trained results can be applied to unseen new videos to support various video-driven facial applications. We further employ a global Multi-Layer Perceptron (MLP) based expression mapping across the latent spaces of various VAEs to preserve characteristics across respective attributes.

Once trained, our neural facial asset can be directly applied to produce unseen facial animations, effectively enriching current scarce dynamic facial assets. Specifically, we present a fast adaptation scheme on the multi-VAE assets to enable cross-identity, fine-grained facial animations driven by in-the-wild video footage. Finally, unlike purely implicit representations, the disentangled facial attributes in our solution support direct geometry and material editing on well-established post-production tools, e.g., for producing stylized character animations. In the following sections, we provide details about our facial data acquisition and neural asset generation in Sec. 3 and Sec. 4, respectively, followed by the performer-specific applications in Sec. 5 (see Fig. 2 for the overview).

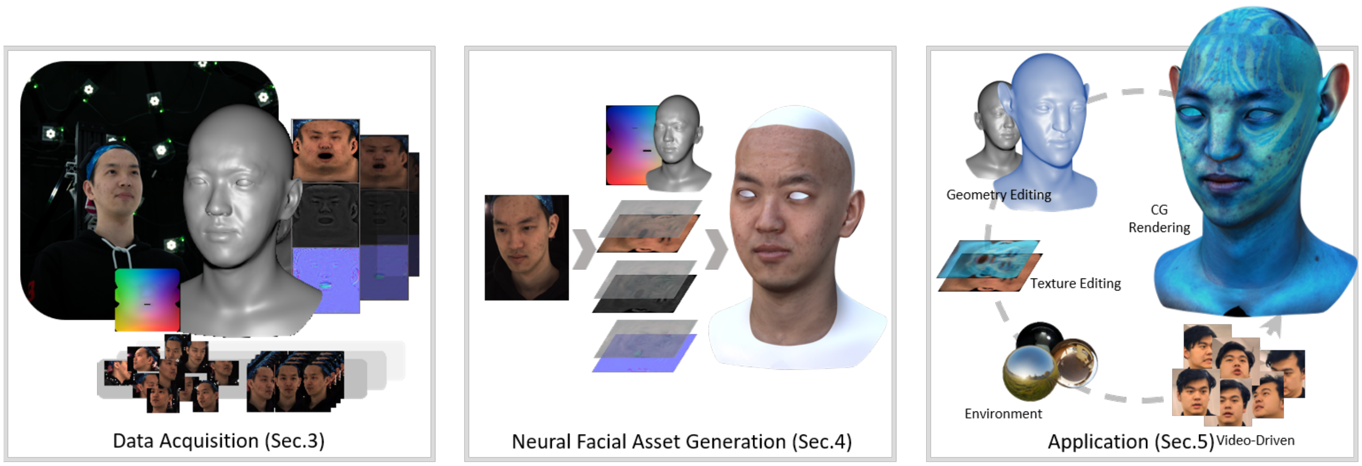



Fig. 2. **The neural physically-based facial asset generation pipeline.** Our data acquisition system, i.e., the Fast Capture Stage (FaStage), extends the classical photometric LightStage [Debevec et al. 2000] by combining multi-view cameras and photometric captures for recovering frame-by-frame detailed facial geometry and physically-based textures at a high frame rate. We then construct a neural representation of the facial asset using specifically designed multi-VAE architecture and training strategies, to simultaneously tackle geometry, textures, and expressions. The resulting neural asset, which is in neural representation, can be driven by in-the-wild face videos with matched facial motion and expression with production-level quality.

## 2 RELATED WORKS

Facial asset generation is a wide field, ranging from high-end productions that rely on sophisticated geometry and motion capture apparatus to low-grade-end consumer applications that use a single camera. For comprehensive reviews, we refer the readers to the recent surveys [Morales et al. 2021; Zollhöfer et al. 2018]. Our work aims to bridge the gap between the two by producing new facial assets from pre-captured ones driven by unseen, single-view videos, at production quality.

*Monocular Facial Reconstruction.* Single-camera-based face animation methods generally adopt parametric face models with adaption for video observation based on landmark detection and head shape regression [Blanz and Vetter 1999; Ekman and Friesen 1978]. Though effective and easy to manipulate, parametric models can not yet reach the highly realistic quality required for production. Blendshape is a prevalent parametric model where artists need to elaborately split or adorn the expressions for photorealistic recreation and animation [Lewis et al. 2014]. It is also possible to register a template model to a real 3D scanned face of the performer under different expressions, e.g., by generating pre-defined blendshapes based on FaceWarehouse [Cao et al. 2014] or 3D morphable models [Egger et al. 2020; Smith et al. 2020]. Weise et al. [2011] fit the dynamic geometry of face to coarse 3D scans from RGBD cameras with pre-defined models whereas Bao et al. [2021] extend the approach for high-fidelity digital avatar generation from RGBD video inputs. Various strategies have been proposed to improve the robustness and accuracy of single image facial asset generation [Cao et al. 2013; Guo et al. 2020, 2018; Paysan et al. 2009; Richardson et al. 2016; Tuan Tran et al. 2017]. Alternative landmark-based reconstructions [King 2009; Zhang et al. 2014a] are epitomized by the FLAME model [Li et al. 2017] that trains on a large dataset of spatial-time 4D scans. To further enhance reconstruction, Cao et al. [2015] set out to add high-fidelity facial details to low-resolution face tracking results. Yet one particular challenge is to reliably handle extreme skin deformation [Chen et al. 2019; Li et al. 2021; Wu et al. 2016] while preserving realism. DECA [Feng et al. 2021] extends FLAME [Li et al. 2017] by predicting performer-specific animatable details that can later be added back to the scanned facial assets of real performers.

To animate facial assets, Fyffe et al. [2014] drive high-quality scans with single-camera inputs whereas Laine et al. [2017] build a learning-based performer-specific framework for production. Various subsequent techniques [Chen et al. 2021; Ma et al. 2021] also support fine-grained performer-specific facial capture. Olszewski et al. [2016] process multiple identities within limited viewpoints whose faces are covered by head-mounted displays. Moser et al. [2021] apply image-to-image translation and extract a common representation for the input video and rendered CG sequence to predict the weights of the target character's PCA blendshapes.

*Physically-based Texture Generation.* In production, physically-based textures such as diffuse albedo, normal maps, specular maps and displacement maps have long served as the key ingredients for creating photo-realistic digital avatars. Ma et al. [2007] pioneer using gradient illumination and diffuse-specular separation with polarizers to acquire high-fidelity physically-based textures. By changing the arrangement of linear polarizers, Ghosh et al. [2011] successfully adapt the previous method to a multiview setting. To support dynamic expressions, Fyffe et al. [2011] use Phantom v640 high-speed cameras to achieve facial performance capture at 264 fps. Fyffe and Debevec [2015] use a color polarized illumination setup to enable texture scanning in a single shot. LeGendre et al. [2018] modify the camera setting to a mono camera to achieve more efficient and higher-resolution results. Riviere et al. [2020] manage to produce high-quality appearance maps from a single exposure. As for full-body performance capture, Guo et al. [2019]

combine spherical gradient illumination with volumetric captures to produce compelling photorealism. With the advances in deep learning, possibilities of estimating the parameters of a predefined reflectance field from single image input [Huynh et al. 2018; Saito et al. 2017; Yamaguchi et al. 2018] have been demonstrated.

Passive facial capture has also made significant progress in recent years, largely attributed to the open-source multi-view reconstruction software [Beeler et al. 2010]. By including skin parameters such as skin color and variations in hemoglobin concentration, Gotardo et al. [2018] manage to obtain more detailed facial features. However, in general, data produced by passive capture methods still cannot match the photometric ones due to calibration errors, camera resolutions, reconstruction errors, etc.

Emerging approaches have employed generative adversarial networks (GANs) to synthesize or infer textures of competent quality for industrial uses. Li et al. [2020a] introduce a framework to generate geometry and physically-based textures from identity and expression latent spaces, with super-resolution networks to produce textures that contain pore-level geometry. Lattas et al. [2021] manage to infer renderable photorealistic 3D faces from a single image. Blendshapes can also be predicted from a single 3D scan of the neutral expression of the performer and then infer dynamic texture maps in a generative manner based on expression offsets [Li et al. 2020b]. Differently, we build dynamic textures using wrinkle maps, which allows us to maintain the authenticity of the high-resolution textures while avoiding the artifacts generated by super-resolution networks.

*Deep Face Models.* It has become increasingly accessible to replace classical geometry and texture-based models with deep models. A neural network can potentially extract a latent vector from the image input and later decode it into high-quality mesh and textures. Chandran et al. [2020] provide a parametric face model with semantic control over the identity and expression of the subject. Lombardi et al. [2018] use a conditional variational autoencoder (CVAE) framework [Kingma and Welling 2014] to encode geometry and view-dependent texture information. Yoon et al. [2019] utilize this model and make it capable of handling monocular input images under uncontrolled lighting environments. Cao et al. [2021] further estimate a gain map and a bias map to additionally improve rendering. [Cao et al. 2022] propose Universal Prior Model (UPM) and produce photo-realistic performer-specific avatars for unseen identities.

GAN-based methods [Gecer et al. 2019; Lattas et al. 2021; Li et al. 2020a] separate the shape and expression or using photorealistic differentiable-rendering-based training to enhance geometry details. Other identity-agnostic methods [Abrevaya et al. 2020; Burkov et al. 2020; Nirkin et al. 2019] aim to enhance the identity-independence by imposing constraints on the latent space. Feng et al. [2018] introduce the UV-position map for 3D geometry representation to enable better regression of CNN. Bi et al. [2021] build a deep face model capable of producing OLAT (one-light-at-a-time) textures, making the predicted avatar relightable under novel lighting environments. In our work, we further improve the quality and editability of the deep face models by generating physically-based textures, owing to the high-quality dynamic facial assets we captured.

# 3 DYNAMIC FACIAL ASSET ACQUISITION

## 3.1 The FaStage

To construct a performer-specific neural facial asset for product-level rendering, the most renowned system, the USC Light Stage [Debevec et al. 2000], exploits the time-multiplexed lighting to recover the facial reflectance field at unprecedented high quality. As shown in Fig. 3, our FaStage adopts a similar semi-dome structure with evenly mounted space-and-time-multiplexed light units. Distinctively, in our FaStage, we replace the DSLR cameras in the USC Light Stage with 3 ultra-fast video cameras and 24 2K video cameras to capture the photos of the subject under 38 gradient illumination patterns and faithfully reconstruct the physically-based dynamic facial performance at video frame rate. We further provide the detailed design and a thorough discussion about our FaStage and captured assets with previous systems [Cao et al. 2014; Cheng et al. 2018; Cosker et al. 2011; Fyffe et al. 2014; Stratou et al. 2011; Yang et al. 2020; Yin et al. 2009, 2008, 2006; Zhang et al. 2014b] in the supplementary material.

## 3.2 Facial Asset Collection

Here we provide the details of collecting topology-consistent assets for typical facial movements at video frame rate, which is critical for exploiting deep networks in our neural facial asset training. Specifically, we capture a total of 5 minutes of facial performance for each performer at 24 fps and acquire the corresponding facial assets. We designed our expression span similar to previous works [Ekman and Friesen 1978; Laine et al. 2017], including rich expressions covering sufficient combinations of facial muscle motions (please refer to the supplementary material).

For these representative performances, we reconstruct the high-quality dynamic facial assets including facial geometry, diffuse albedo, specular intensity, and normal map from the data captured by our FaStage. During capture, we send one specific control signal to turn all the lights on and collect the full-lit images from both the machine vision and high-speed cameras. Then, we perform the multi-view stereo technique on full-lit images to extract the facial geometry and fit a human face model [Li et al. 2020a] to obtain topology-aligned results similar to previous work [Cosker et al. 2011]. We also follow the work of Li et al. [2020a] to unwarp the facial geometry into UV maps and uniformly sample 256×256 points where each pixel represents the location of the corresponding point on the mesh to generate geometry maps. To obtain physically-based textures, we apply 38 gradient illumination patterns for every 1/24 second and then use the 38 images from high-speed cameras under the corresponding patterns to recover diffuse albedo, specular intensity, and object-space normal map at 24 fps using the photometric stereo technique [Ma et al. 2007]. We adopt the standard fusion method [Ghosh et al. 2011] to merge the corresponding texture across different views in UV space and further transform the object-space normal maps into tangent space to reduce temporal noise.

*Temporal Stabilization.* We recognize the noticeable temporal jittery artifacts on the captured physically-based textures after the above unwrapping and multi-view merging process, especially for

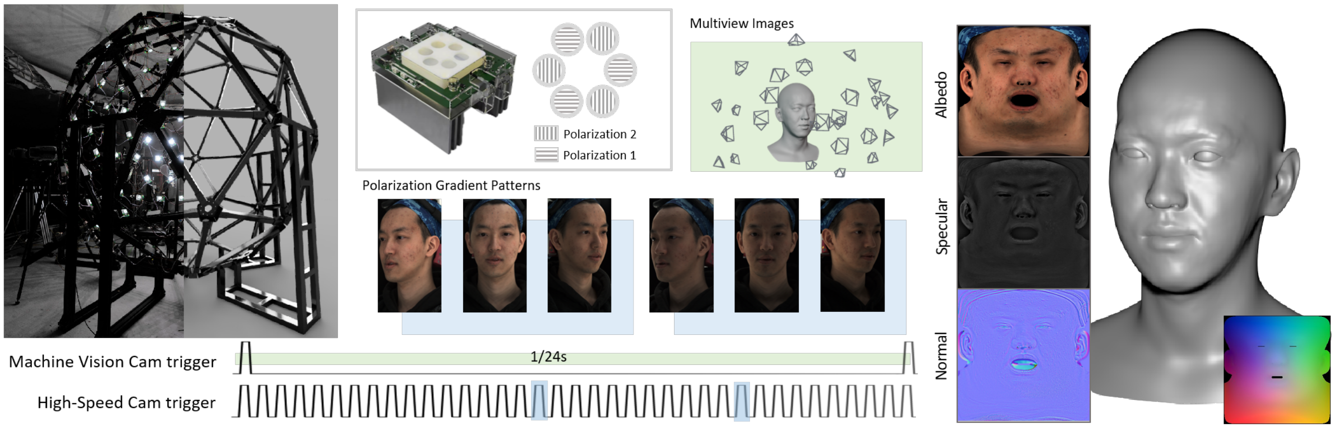



Fig. 3. **Illustration of our data acquisition pipeline.** FaStage uses three high-speed cameras to capture polarization gradient patterns for dynamic physically-based texture acquisition. Multiple machine vision cameras are synchronized at a low frequency for dynamic facial geometry acquisition.

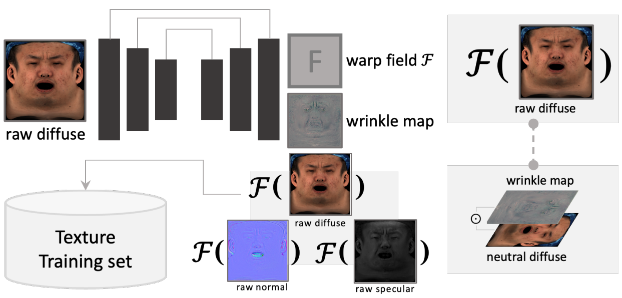



Fig. 4. **Pipeline of texture stabilization.** We utilize an unsupervised method to predict the warp field with regard to the neutral textures for textures of each frame. Stabilized textures, which are produced by applying the warp field to the textures of each frame, will be used as training data.

our 4D capture setting at video frame rate, leading to unreliable neural asset training. Thus, we introduce a novel self-supervised scheme to stabilize the obtained temporal textures, as illustrated in Fig. 4. We adopt a U-Net structure to predict the temporal warp fields and wrinkle maps for the original textures in each frame with regard to the neutral one. Since the subtle appearance details in diffuse albedo are more suitable for temporal alignment, we only adopt the diffuse albedo for our stabilization scheme. To this end, given an input diffuse albedo $\mathbf{T}_t$ at $t$-th frame, our stabilization network $\mathbf{\Phi}$ predicts both a warp field $\mathcal{F}_t$ and a corresponding wrinkle map $\mathcal{W}_t$:

$$\mathcal{F}_t, \mathcal{W}_t = \Phi(\mathbf{T}_t). \tag{1}$$

Here, $\mathcal{F}_t$ represents pixel-wise flow and we use $\mathcal{F}_t(\cdot)$ to represent the warping of a texture to the neutral frame where the performer makes no expressions. The wrinkle map $\mathcal{W}_t$ denotes the element-wise multiplication maps on top of the neutral texture in the UV domain to represent the stretch and squeeze of wrinkles from neutral to current expression caused by facial motions. It ranges from 0 to 1 to maintain the original brightness of captured diffuse albedo. Then, we formulate the following fidelity loss to ensure temporally aligned diffuse albedo after applying the warp field $\mathcal{F}_t$ to $\mathbf{T}_t$:

$$\mathcal{L}_{fid} = \|\mathcal{F}_t(\mathbf{T}_t) - \mathcal{W}_t \odot \mathbf{N}\|_1, \tag{2}$$

where $\mathbf{N}$ is the neutral diffuse albedo and $\odot$ is element-wise multiplication. We also adopt a temporal consistency term to ensure texture stability between consecutive frames to implicitly align two textures. Given the network output $\mathcal{W}_t, \mathcal{W}_{t+1}$ from consecutive frames, the temporal term is formulated as:

$$\mathcal{L}_{tem} = \|\mathcal{W}_t - \mathcal{W}_{t+1}\|_1. \tag{3}$$

In addition, we apply a regularization term to restrict and regularize the estimated flow and its gradient:

$$\mathcal{L}_{dis} = \|\mathcal{F}_t\|_2 + \|\nabla \mathcal{F}_t\|_2. \tag{4}$$

Our total loss for texture stabilization is formulated as follows:

$$\mathcal{L}_{stab} = \mathcal{L}_{fid} + \mathcal{L}_{tem} + \mathcal{L}_{dis}, \tag{5}$$

where weights for balancing different losses are omitted for the simplicity of presentation. Note that the diffuse albedo, specular intensity, and normal map share the same warp field, therefore we apply $\mathcal{F}$ to all the other texture attributes similar to the diffuse albedo. For training our network $\Phi$, we use all the captured dynamic textures to optimize $\mathcal{L}_{stab}$. We downsample the input textures to 512×512 to predict the corresponding warp fields and wrinkle maps. We implement our U-Net architecture and perform the training process according to Pytorch-UNet [milesial 2022]. Once trained, we apply the warp fields from $\Phi$ to process all the textures to obtain stabilized ones for the generation of neural facial assets in Sec. 4.

## 4 NEURAL FACIAL ASSET GENERATION

From captured dynamic facial assets of a specific performer, we intend to transform them into a drivable neural representation. Traditional human facial performance capture relies on heavy manual labor to align geometry and textures, rigging, and expression deformation. This process can be potentially bypassed by a neural-based approach. To this end, we build a learning framework with three networks for disentangling the expression, facial geometry and

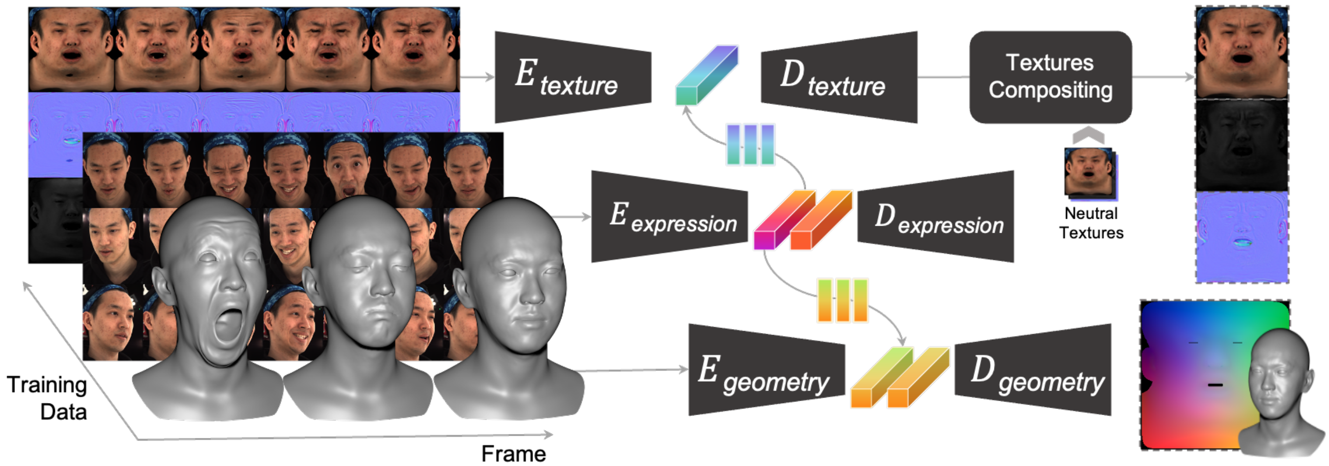



Fig. 5. **Neural physically-based facial asset.** For each frame, the training input consists of multi-view facial images captured in FaStage, as well as the coarse geometry and physically-based textures. Our approach trains three deep Variational Auto-Encoders to obtain three latent spaces separately on captured image, geometry and textures. Then two multilayer perceptrons are trained for latent space translation from expression to geometry and textures.

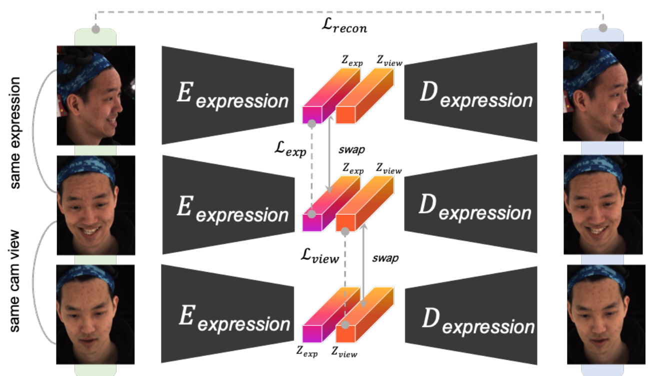



Fig. 6. **Architecture of expression network.** We constrain the encoder to disentangle expression and view.

physically-based texture. The whole training pipeline is illustrated in Fig. 5.

## 4.1 Expression and Viewpoint Disentanglement

We train a VAE to learn a global representation for expressions and viewpoints from training footage, as illustrated in Fig. 6. Given an input image $I$ from view $i$ at frame $t$, our expression encoder $\mathcal{E}_e$ encodes it to an expression latent code $Z_e$ and a viewpoint latent code $Z_v$, while our expression decoder $\mathcal{D}_e$ predicts the reconstructed image $\hat{I}$ given these two latent codes. To ensure latent space consistency between the encoder and decoder, we predict cycled latent codes $\hat{Z}_e, \hat{Z}_v, (\hat{Z} = \mathcal{E}[\mathcal{D}(Z)])$, from reconstructed image $\hat{I}$, and adopt the reconstruction and latent consistency loss:

$$\mathcal{L}_{rec} = \|I - \hat{I}\|_1, \;\; \mathcal{L}_{cyc} = \|Z_e - \hat{Z}_e\|_2 + \|Z_v - \hat{Z}_v\|_2. \tag{6}$$

Notice we omit view and frame index $i$ and $t$ in the equation for simplicity in presentation.

*Triplet supervision for expression encoding.* We designed a triplet training scheme which applies a latent swapping method [Aberman et al. 2019; Huang et al. 2018; Park et al. 2020] that focuses on retrieving consistent expression and viewpoint information across views and frames. In each training iteration, we randomly pick two frames $s, t$, and two random views $i, j$. Notice all frames and views data are from the same performer. Given the input images $I^{t,i}$, $I^{t,j}$, $I^{s,i}$, our encoder $\mathcal{E}_e$ predicts the performer's expressions and the camera's viewpoint latent codes for each of them, i.e., $Z^{t,i}, Z^{t,j}, Z^{s,i}$: Notice the performer exhibits the same expression in $I^{s,i}$ and $I^{s,j}$, while in $I^{s,i}$ and $I^{t,i}$, the viewpoints are the same. Therefore, we minimize the distance between the expression latent code and the view latent code to achieve expression-viewpoint disentanglement:

$$\mathcal{L}_{exp} = \|Z_e^{t,i} - Z_e^{t,j}\|_2, \;\; \mathcal{L}_{view} = \|Z_v^{t,i} - Z_v^{s,i}\|_2. \tag{7}$$

Besides constraining the encoder, we want the decoder to focus on recovering the expression despite the disturbances from viewpoint changes. Therefore, the subsequent reconstruction is done in a cross frame and view manner:

$$\begin{aligned} \tilde{I}^{t,i} &= \mathcal{D}_e(Z_e^{t,j}, Z_v^{s,i}), \\ \tilde{I}^{t,j} &= \mathcal{D}_e(Z_e^{t,i}, Z_v^{t,j}), \\ \tilde{I}^{s,i} &= \mathcal{D}_e(Z_e^{s,i}, Z_v^{t,i}). \end{aligned} \tag{8}$$

The decoded result $\tilde{I}^{t,i}$ should be the same as the original image $I^{t,i}$, since the encoded expression and viewpoint information are supposed to be the same. The corresponding reconstruction error at this stage is formulated as:

$$\mathcal{L}_{cro} = \|I^{t,i} - \tilde{I}^{t,i}\|_1 + \|I^{t,j} - \tilde{I}^{t,j}\|_1 + \|I^{s,i} - \tilde{I}^{s,i}\|_1. \tag{9}$$

Then the total loss for our expression network can be formulated as follows:

$$\mathcal{L}_E = \mathcal{L}_{rec} + \mathcal{L}_{cyc} + \mathcal{L}_{exp} + \mathcal{L}_{view} + \mathcal{L}_{cro}. \tag{10}$$

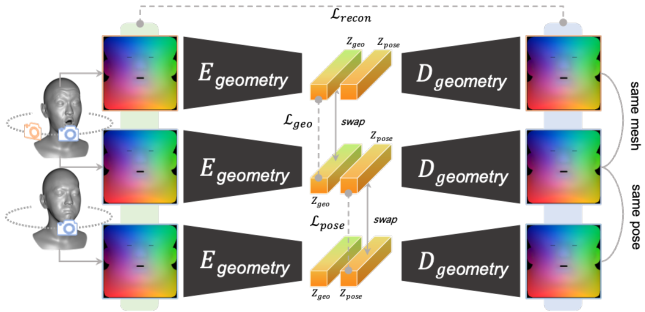



Fig. 7. **Architecture of geometry network.** We constrain the encoder to disentangle geometry and pose.

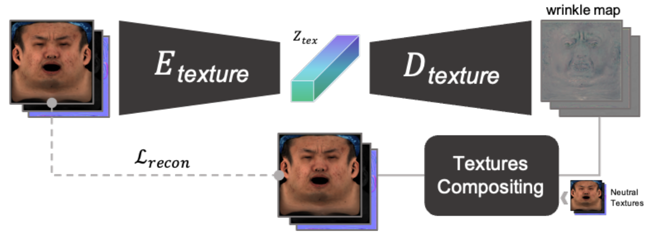



Fig. 8. **Architecture of texture network.** To enable high-resolution dynamic textures, our texture network outputs wrinkle maps instead of textures themselves.

This loss helps us to constraint $Z_e$ so that it only contains the expression features, and the viewpoint information is in $Z_v$. Notice we omit balancing weights for simplicity in presentation.

## 4.2 Geometry Extraction

Recall that we represent the facial geometry with a 2D-geometry map in UV space where each pixel represents the 3D coordinate of the corresponding vertex. Therefore, we can regard geometry inference as an image reconstruction task. To extract better geometry features and eliminate the minor head pose errors, we adopt a similar network structure to the one developed for expression encoding. For improving robustness, we apply random rotations and translations to the scanned geometries for data augmentation. As illustrated in Fig. 7, given an input geometry map $G$ from frame $t$ with augmentation operation $a$, our geometry encoder $\mathcal{E}_g$ encodes it to a geometry latent code $Z_g^{t,a}$ and a pose latent code $Z_p^{t,a}$, while our geometry decoder $\mathcal{D}_g$ predicts the reconstructed image $\hat{G}^{t,a}$ given these two latent codes. Similar to the expression disentanglement, we randomly pick frames with $s, t$ and augmentation operations $a, b$ for training. Note that we choose $s, t$ from the same dynamic clip and make sure they are temporally close to ensure both geometries without augmentation have similar poses. Specifically, when we fix the geometry latent code and change the pose latent code, the geometry maps generated by the decoder should exhibit similar geometry and only differ in rotation and translation. Therefore, the decoding process itself is designed to be independent of the pose latent code, which could be formulated as:

$$\mathcal{D}_g(Z_g, Z_p) = \mathbf{M}_p(Z_p) \otimes \mathcal{D}^{\dagger}{}_g(Z_g), \tag{11}$$

where $\mathbf{M}_p$ is a multilayer perceptron (MLP) for transforming the pose latent code with a rotation and translation matrix, $\mathbf{D}^{\dagger}{}_g$ represents the network only for decoding the geometry latent code, and $\otimes$ represents matrix-vector multiplication. Notice we omit the frame index $t$ and the augmentation operation $a$ for simplicity. In this way, our network is able to predict the performer's facial geometry free from the disturbance of head pose variance.

We use the same triplet supervision strategy as the expression network to disentangle the facial geometry and head pose, which has the following metrics:

$$\begin{aligned}
&\mathcal{L}_{rec} = \|G^{t,a} - \hat{G}^{t,a}\|_2, \\
&\mathcal{L}_{cyc} = \|Z_g^{t,a} - \hat{Z}_g^{t,a}\|_2 + \|Z_p^{t,a} - \hat{Z}_p^{t,a}\|_2, \\
&\mathcal{L}_{geo} = \|Z_g^{t,a} - Z_g^{t,b}\|_2, \\
&\mathcal{L}_{pos} = \|Z_p^{t,a} - Z_p^{s,a}\|_2, \\
&\mathcal{L}_{cro} = \|G^{t,a} - \tilde{G}^{t,a}\|_2 + \|G^{t,b} - \tilde{G}^{t,b}\|_2 + \|G^{s,a} - \tilde{G}^{s,a}\|_2.
\end{aligned} \tag{12}$$

Where $\mathcal{L}_{rec}$, $\mathcal{L}_{cyc}$ are reconstruction and cycled latent loss. $\mathcal{L}_{geo}$ and $\mathcal{L}_{pos}$ are the distances between latent codes. $\mathcal{L}_{cro}$ is the cross inference loss, similar to the definition in Eq. 9. The total loss for the geometry network is formulated as follows:

$$\mathcal{L}_G = \mathcal{L}_{rec} + \mathcal{L}_{cyc} + \mathcal{L}_{geo} + \mathcal{L}_{pos} + \mathcal{L}_{cro}. \tag{13}$$

## 4.3 Learning Textures

The stabilized textures are view-independent and only contain expression information. Therefore, we build a similar VAE with only one latent space to learn the representations for the three types of textures (diffuse albedo, specular intensity, and normal maps) simultaneously. The network structure is illustrated in Fig. 8. Notice the captured texture is at 4K resolution, yet the output resolution of the network is typically 512, which inevitably leads to a significant loss of details. To retain the ability of rendering and editing at 4K resolution, we enforce this network to predict wrinkle maps (we borrow this term from the CG industry ) instead of predicting the textures directly. Given an input texture $T$ at frame $t$, our texture encoder $\mathcal{E}_{tex}$ encodes it to a texture latent code $Z_{tex}$, while our texture decoder $\mathcal{D}_{tex}$ predicts the corresponding wrinkle maps $\hat{W}$ from the texture latent code $Z_{tex}$:

$$\hat{W} = \mathcal{D}_{tex}(Z_{tex}), \;\; \hat{T} = \hat{W} \oplus \mathbf{N}, \tag{14}$$

where $\mathbf{N}$ is the neutral texture, $\hat{T}$ represents the reconstructed textures, and $\oplus$ represents the addition operation. Similarly, we compute the texture reconstruction loss and cycled latent loss. During the training process, the losses are formulated as:

$$\mathcal{L}_{rec} = \|T - \hat{T}\|_1, \;\; \mathcal{L}_{cyc} = \|Z_{tex} - \hat{Z}_{tex}\|_2, \tag{15}$$

and the total loss for the texture network is formulated as follows:

$$\mathcal{L}_T = \mathcal{L}_{rec} + \mathcal{L}_{cyc}. \tag{16}$$

### 4.4 Unified Expression Representation

For each training frame $t$, our previously discussed three networks predict the expression latent code $Z_e$, the geometry latent code $Z_g$ and the texture latent code $Z_{tex}$ respectively. Intuitively, we can learn a mapping from expressions to geometries and textures, as we have disentangled head pose and viewpoint into separate latent codes. We train an MLP $\mathbf{M}_e$ to learn mappings from the facial expression latent code to the geometry and texture latent codes. Given an expression $Z_e$ as input, $\mathbf{M}_e$ directly predicts $\tilde{Z}_g$ and $\tilde{Z}_{tex}$ as:

$$\tilde{Z}_g, \tilde{Z}_{tex} = \mathbf{M}_e(Z_e). \tag{17}$$

The mapping loss is straightforward which minimizes the distance between the encoded and mapped latent codes:

$$\mathcal{L}_{map} = \|Z_g - \tilde{Z}_g\|_2 + \|Z_{tex} - \tilde{Z}_{tex}\|_2. \tag{18}$$

### 4.5 Training and Inference

*Training.* Here we introduce an effective training scheme for our neural representation. We first separately train our expression, geometry and texture VAE branches (Sec. 4.1, 4.2, 4.3 respectively), to learn latent representations independently. After the three networks converge, we fix VAEs in all three branches and only train the latent mapping MLP $\mathbf{M}_e$. Then, we apply an additional fine-tuning process to optimize the total loss. In this fine-tuning stage, we only lock the parameters of the expression network, and jointly train all other parts of the whole network (i.e., the geometry, texture, and mapping network). This strategy keeps the expression latent code stable and allows for more precise geometry and texture adjustment by fine-tuning the geometry and texture latent representations. Note that in the actual implementation, we minimize the Kullback-Leibler (KL) divergence loss for the latent codes in VAEs as well.

*Inference.* Our neural facial asset representation unifies the person's specific expression, geometry, and texture latent codes. It grants us the ability to decode the geometry and physically-based texture assets directly from expressions. At the inference stage, the input is a face image with certain expressions, we then generate the expression latent code using $\mathcal{E}_e$ and use $\mathbf{M}_e$ to translate the encoded expression latent code to geometry and texture latent codes. Subsequently, we use the geometry and texture decoders $\mathcal{D}_g$ and $\mathcal{D}_{tex}$ to generate the facial geometry, albedo, specular map, and normal map. We formulate the inference process as: given an input image $\mathbf{I}$, we predict both the geometry $\tilde{\mathbf{G}}$ and the textures $\tilde{\mathbf{T}}$ with our expression encoder, latent mapping multilayer perceptron, and geometry and texture decoder:

$$\begin{aligned} Z_e &= \mathcal{E}_e(I), \\ \hat{Z}_g, \hat{Z}_{tex} &= \mathbf{M}_e(Z_e), \\ \tilde{G} &= \mathcal{D}_g(\hat{Z}_g), \\ \tilde{W} &= \mathcal{D}_{tex}(\hat{Z}_{tex}), \\ \tilde{T} &= \tilde{W} \oplus \mathbf{N}, \end{aligned} \tag{19}$$

where the viewpoint latent code and the geometry pose latent code are ignored since we intend to produce results that only depend on the facial expressions of the performer. Such pipeline produces consistent geometry and textures across different views of the input images due to our latent disentanglement network design.

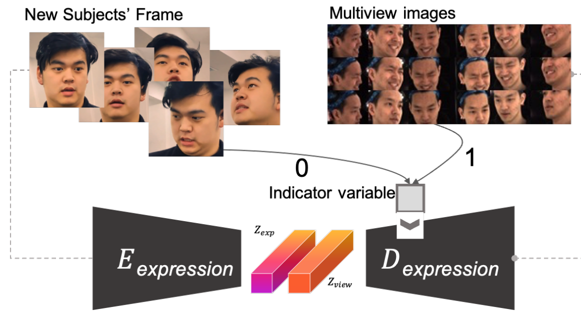



Fig. 9. **Indicator variable.** The indicator variable informs the decoder of which domain it should decode the image into.

We provide the implementation details of our framework in our supplementary material. For further application of cross-identity neural retargeting, which will be mentioned in Sec. 5.2, inspired by Lombardi et al. [2018], we propose a similar but different network design. In our expression decoder, we concatenate a channel of zeros to each input before convolutional layers, denoted as **indicator channel**, which will be further used for cross-identity neural retargeting in Sec. 5.2.

## 5 APPLICATIONS

Once our neural facial asset generator is trained for a specific performer, we can apply it in several scenarios with slight modification of our network. We can generalize our network to predict the performer's appearances with novel expressions, even animated by another person's in-the-wild face video. We also introduce various editing effects using our facial assets, including material editing, character animation and wrinkle transfer.

### 5.1 Performer-specific neural facial asset

Once trained, our neural facial asset can be directly applied to produce unseen facial animations of the performer. It takes the monocular camera footage as input and outputs the geometry and corresponding multi-channel wrinkle maps, producing photo-realistic rendering results under novel facial animations from novel views and lighting. Thus, it can effectively enrich current scarce dynamic facial assets, and makes it possible to preview the dynamic facial assets instantly. Such strategy removes the process to capture and reconstruct unseen facial animations using the complex hardware setup.

### 5.2 Cross-identity neural retargeting

The movie and game industry has always had a keen interest in facial performance capture. A common request from the industry is to drive the face animation of a digital character from the facial expressions of a real performer. Recall that our expression encoder can extract the performer's expressions, hence it only needs minor refinement when we apply it to another individual, as presented in Fig. 9.

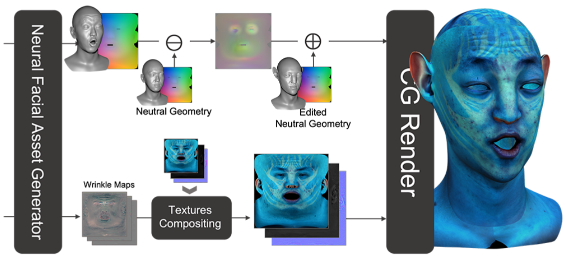



Fig. 10. **Illustration of geometry and texture editing pipeline.** We can produce the realistic appearance of a novel character with consistent expressions.

Specifically, given a well-trained neural physically-based facial asset, first, we extend the expression training dataset with footage of the new subject with different expressions and varying head poses. Then, we refine the expression network by training on this extended dataset. Besides the losses mentioned in Eq. 10, we simultaneously minimize the reconstruction loss of images from the new subject with a strategy similar to the ones used in [Higgins et al. 2016] and [Lombardi et al. 2018]. To let the expression decoder reconstruct images for the driving subject, we set the **indicator channel** of the decoder to 1 for images from footage of the driving subject and 0 for images from the original dataset. This indicator allows the expression latent code to contain no identity-specific information but only expressions. Moreover, to ensure the encoder produces the expression latent code independent from identity, we further add a discriminator with RevGrad layer Ganin and Lempitsky [2015] to supervise the expression latent code.

Once trained the modified expression network, we lock its parameters and then refine both the geometry network and texture network together with the latent mapping MLP $\mathbf{M}_e$ on the original training dataset. Then, the new subject's performance can be transferred to the original performer. The disentanglement of expression and viewpoint also allows the encoder to stably work with images taken under varying head poses, which means the new subject can even drive the performer with portable capture devices, such as cellphones.

## 5.3 Geometry and texture editing

Animation of computer augmented faces is essential for the movie and game industry, which involves geometry and texture editing of the captured face model and requires artist labor. For instance, animating digital characters in the Avatar movie involves manual creation of novel appearances and hundreds of FACS action units and corresponding textures in order to generate realistic appearances, expressions and wrinkles. Previous works focus on high-precision facial capture or neural rendered final effects, yet they can not provide the artists with sufficient editability in textures, geometry, and shader to release the manual editing burden. In contrast, with our method, driving a stylized character becomes much easier, where only the neutral geometry and textures for the fictional character are needed. The editing effort can be significantly reduced, as presented in Figure 10.

To drive the stylized character, we first create the character under neutral expression with artificial facial details and novel skin appearances. We can modify the geometry to change facial features, e.g. longer ears and thinner cheeks. Similarly, we can modify the textures by painting on diffuse albedo, changing micro-geometry on normal map and editing specular intensity. We expect this new character to exhibit similar movements in facial muscle and wrinkle details. Then, we transfer the dynamic facial geometry as well as wrinkle maps under different expressions to drive this new character.

Taking the driving subject's facial performance video as input, our neural facial asset can predict per-frame dynamic geometry. We calculate the facial movement, which we regard as an offset field, and then apply the offset to the neutral geometry map of stylized character to obtain the geometry with the same expression. During the texture compositing stage, we replace the original neutral textures with an exquisitely drawn one. Because our network is designed to predict wrinkle maps, it can transfer the performer's wrinkles to the stylized character.

## 6 RESULTS

In this section, we demonstrate the capability of our approach in a variety of scenarios. We first report the implementation details of our approach and the dataset captured by our FaStage, and then analyze our results in various editing scenarios. We further provide a comparison with previous methods and evaluations of our main technical components, both qualitatively and quantitatively. The limitation and discussions regarding our approach are provided in the last subsection.

*Dataset and Implementation Details.* To train and evaluate our method, we capture a dynamic facial asset dataset with physically-based textures using our FaStage system. Our capture system is shown in Fig. 3, where all the cameras are calibrated and synchronized with lighting in advance. During capture, the machine vision camera array produces 24 RGB streams at 2592×2048 resolution, while the high-speed camera array generates 3 RGB streams at 4096×2304 resolution. After data pre-processing, we produce 4096x4096 unwrapped physically-based textures for each frame, including diffuse albedo, normal map, and specular intensity. Our dataset consists of three performers and we capture a total number of 7200 frames for each performer, respectively. In our implementation of facial motion re-targeting, we use the first 30 seconds of the in-the-wild video clip for refinement. We crop the frames in the clip to 256x256 with faces in the center using the landmark detector from Dlib [King 2009], and only compute reconstruction loss within the face region. Additionally, we adopt Blender's Eevee as our main render engine as it is free, open-source, fast, and convenient. Rendering a frame with textures of 4K resolution takes less than 1 second with the Eevee engine. We further utilize the rendering shader provided in the Heretic [Unity Technologies 2020] to achieve real-time performance. Fig. 11 demonstrates several neural facial assets generated using our approach for high-realistic facial capture and physically-based editing.

Fig. 11. **Driving and re-targeting of facial motions from videos using our neural physically-based facial assets.** Each tuple includes driving frames and facial assets, with style of the actor himself (left), Guan Yu (middle), and Alita (right).

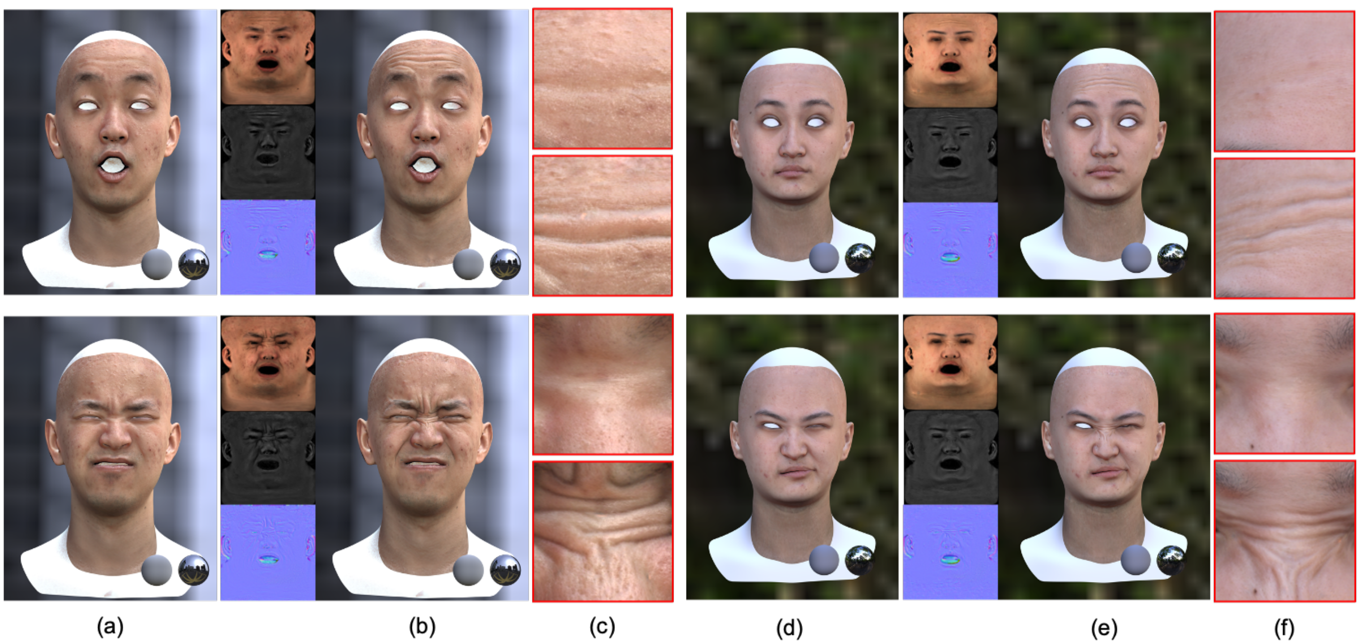



Fig. 12. **Video-driven Neural Physically-based Facial Asset.** Our dynamic textures effectively enhance the appearance of the performer: (a,d) driven facial assets with only static neutral physically-based textures, (b,e) driven facial assets with dynamic physically-based textures, (c,f) zoom-in view. Our method successfully models the dynamic textures and preserves facial details at high resolution.

### 6.1 Video-driven Applications on Novel Expressions

Our multi-VAE-based neural facial assets enable various video-driven applications, ranging from fine-grained facial capture to cross-identity facial motion retargeting, and to geometry and physically-based texture editing. As shown in Fig. 12, our approach enables high-quality dynamic facial geometry generation with pore-level dynamic textures on complicated expressions of the specific performer, such as frowning or winking. Fig. 13 further provides the video-driven results using the RGB footage of different performers. In addition to driving the captured performer, our method further enables convenient physically-based editing for generating and driving an artistically stylized avatar, as shown in Fig. 14. Besides, our retargeted animations are temporally consistent, which is provided in the accompanying video.

### 6.2 Comparisons

To demonstrate the video-driven performance of our approach, we compare against various video-driven facial reconstruction and animation methods, including parametric facial model 3DDFA proposed by Guo et al. [2020] and DECA proposed by Feng et al. [2021], a performer-specific method for production proposed by Laine et al.

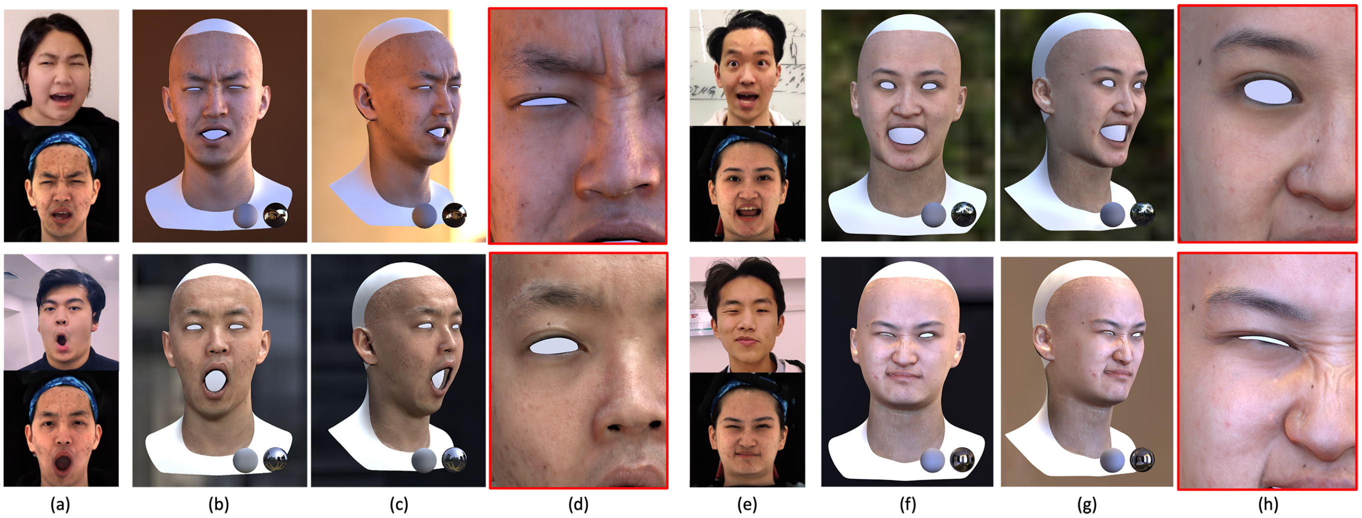



Fig. 13. **Cross-identity video driven results.** We extend the expression network for new subjects with different expressions and varying head poses: (a,e) the upper image is the new performer's expression, and the lower image is the neural retargeting results from our expression decoder, (b,f) driven facial asset in front view, (c,g) driven facial asset in the left view, (d,h) zoom-in view. Our method achieves detailed video-driven results from different identities with dynamic textures, which leads to photo-realistic rendering.

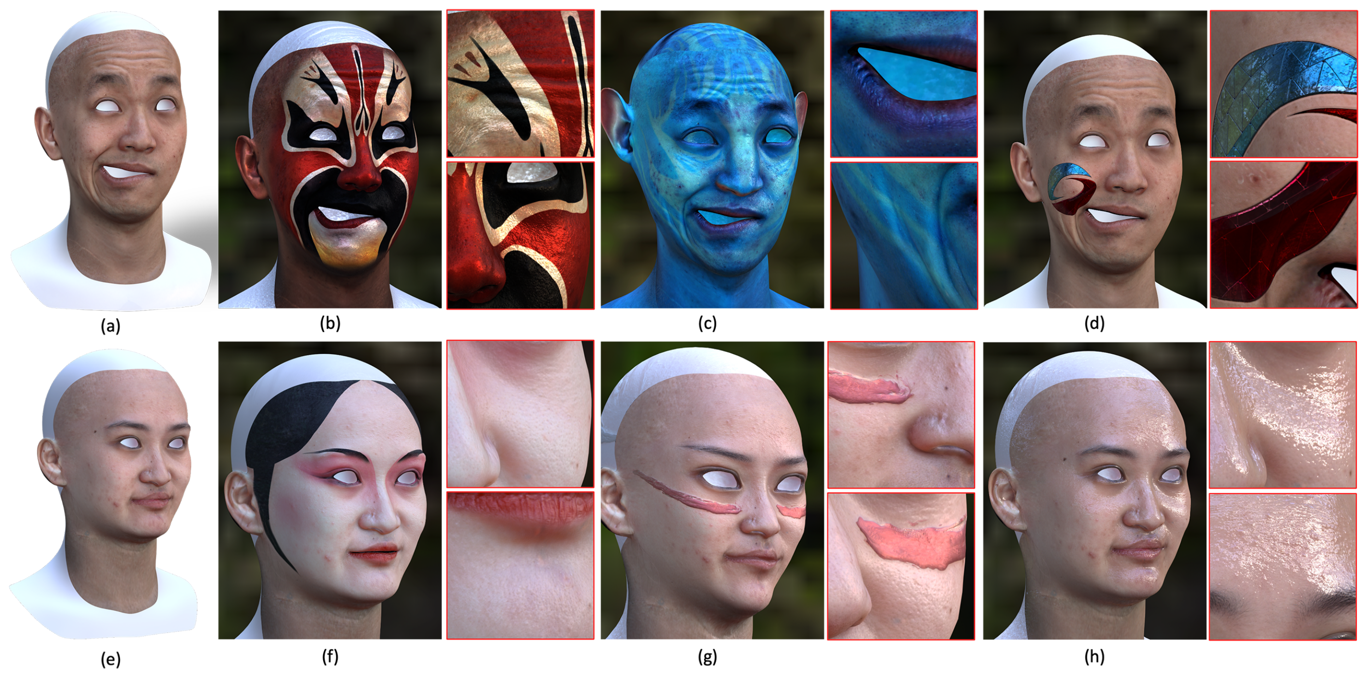



Fig. 14. **Geometry and texture editing results.** We edit the video-driven facial assets (a,e) in various ways: (b) We paint the diffuse albedo with the style of Guan Yu, the most famous red-face character in traditional Chinese Peking Opera; (c) We stylize the geometry of the performer to the blue alien from the feature film AVATAR with a novel facial structure but consistent identity features; (d) We add a metal SIGGRAPH logo to the cheek of the performer by modifying the textures; (f) We paint the diffuse albedo with the style of Yu Ji, the beloved concubine of Xiang Yu, the hegemon of Western Chu, from the famous Peking Opera *Farewell My Concubine*; (g) We alter the performer's facial structure, and added realistic face painting by jointly modifying the diffuse albedo and the normal map; (g) We adjust the facial roughness to give the skin a more shiny look.

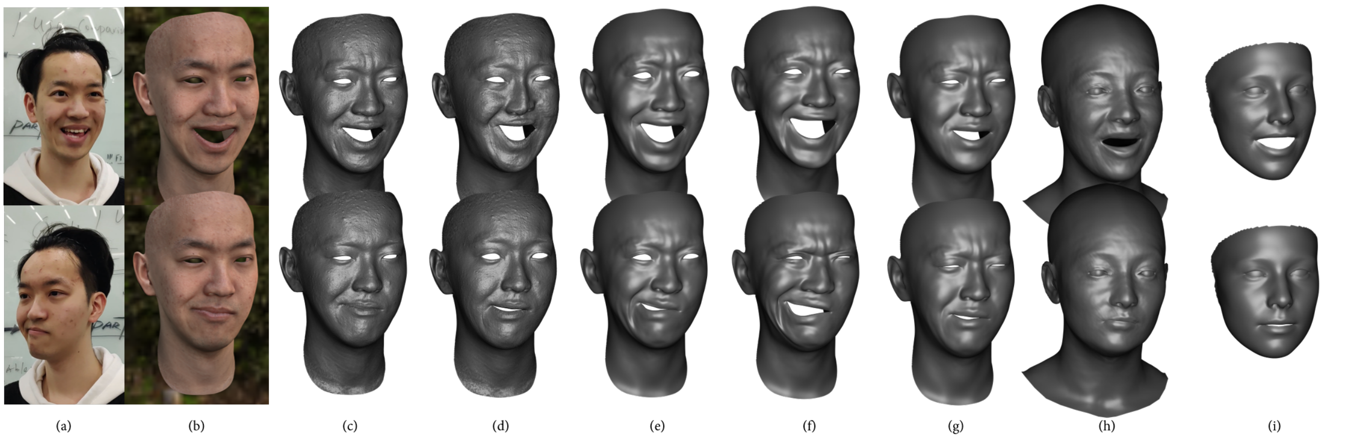
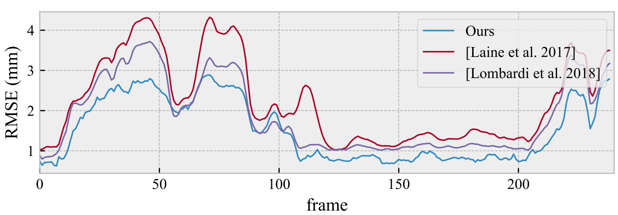



Fig. 15. **Qualitative comparison of video-driven result on in-the-wild monocular RGB inputs.** (a) the input in-the-wild image, (b) rendered result of our method with textures, (c) predicted geometry of our method with normal map, after the refinement using in-the-wild images, (d) predicted geometry of our method with normal map, before the refinement using in-the-wild images, (e) predicted geometry of Lombardi et al. [2018], trained with in-the-wild images, (f) predicted geometry of Lombardi et al. [2018], trained without in-the-wild images, (g) predicted geometry of Laine et al. [2017], (h) predicted geometry of Feng et al. [2021], (i) predicted geometry of Guo et al. [2020]. With a quick refinement, our method can deal with head pose and lighting variance better than other methods.




| Method | RMSE (mm) | | |
|---|---|---|---|
| | total | eye | mouth |
| Laine et al. [2017] | 2.237 | 1.565 | 4.157 |
| Lombardi et al. [2018] | 1.866 | 1.174 | 3.076 |
| **Ours** | **1.516** | **1.164** | **2.750** |

Fig. 16. **Quantitative comparison against several state-of-the-art methods.** We compare the per-frame geometric reconstruction errors on an image sequence in our testing dataset (upper) and overall geometric reconstruction errors on the images from our testing dataset (lower). Our method outperforms others, indicating the better expression ability of our network.

[2017], and Deep Appearance Model proposed by Lombardi et al. [2018]. We adopt the official pre-trained PyTorch model for the method of Feng et al. [2021] and Guo et al. [2020], and faithfully re-implement both the methods of Laine et al. [2017] and Lombardi et al. [2018] and train their models using the same training dataset as ours for a fair comparison.

We divide 20% of the total frames of an individual as our testing set. We provide multi-view images as input for both the methods of Laine et al. [2017] and Lombardi et al. [2018] in the training process. Fig. 15 provides the qualitative comparison of our testing dataset for the task of facial geometry capture from in-the-wild monocular video input. It is worth noting that compared to other methods, our approach generates more realistic pore-level facial details which are exhibited in normal maps. Besides, our approach generates more realistic video-driven facial animation results even under challenging expressions, which compares favorably to other methods. For fair quantitative comparison, we first transform the output models on our testing dataset from both Laine et al. [2017] and Lombardi et al. [2018] into our geometry map representation. Note that the unit of coordinates is transformed into millimeters and we adopt RMSE as the metric. As shown in Fig. 16, our approach consistently outperforms the baselines, which demonstrates the effectiveness of our approach for dynamic video-driven facial capture and modeling. More importantly, our approach enables various physically-based applications like geometry and material editing or facial feature transfer, which are rare in the above methods.

## 6.3 Ablation Study

Here, we evaluate the performance of different training strategies and demonstrate the effectiveness of our multi-VAE network designs. Specifically, we evaluate the temporal stabilization, the expression disentanglement, the pose enhancement, and wrinkle-based texture prediction in our expression, geometry, and texture VAE, respectively.

*Temporal stabilization.* We compare the stability of our textures with or without the temporal stabilization. Let **w/o stabilization** denote the results of our original textures without temporal stabilization. We compare the difference of textures between consecutive frames, where the lower difference indicates less temporal instability. As illustrated in Fig. 17, our temporal stabilization significantly improves the stability of textures for later network training.

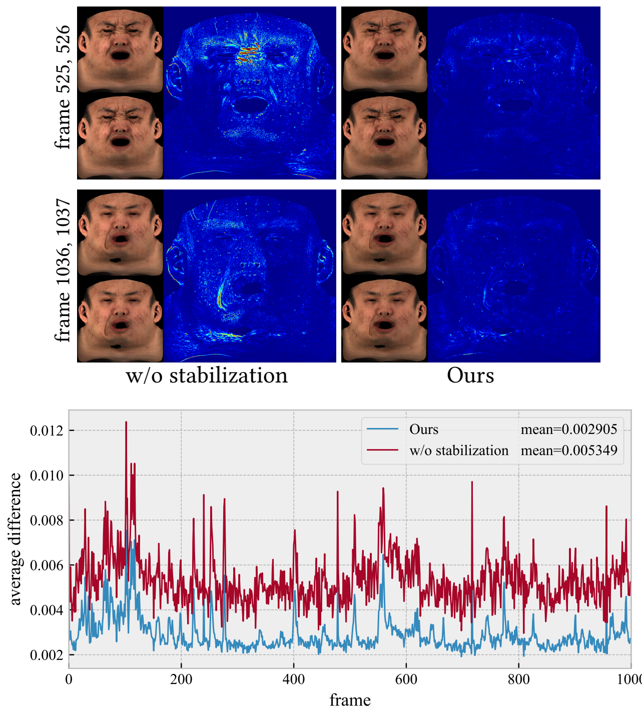



Fig. 17. **Effectiveness of temporal stabilization.** The upper figures show the per-pixel difference of diffuse textures from consecutive frames with and without temporal stabilization. The lower curve shows the average difference of consecutive diffuse textures in our first 1000 frames. With temporal stabilization applied, we significantly reduce the difference between consecutive frames and provide more consistent textures.

*Expression disentanglement.* We compare the performance of our model with or without the expression disentanglement training scheme. Let **w/o disentanglement** denote the variation of our approach without triplet input for both expression and geometry VAEs. We utilize the same multi-view images as input for training but ignore the viewpoint information. During the training process of the expression network, we only adopt the loss $\mathcal{L}_{rec}$ in Eq. 6, and remove viewpoint latent code from our network. As shown in Fig. 18, our scheme with expression disentanglement enables more accurate and consistent prediction for geometry, especially when the input viewpoint changes. Our expression disentanglement training scheme facilitates the robustness of our model to handle viewpoint variance.

*Pose extraction.* Here we compare against our variation without encoding the pose latent code from geometry while keeping the input geometry intact without augmentation, denoted as **w/o pose**. During the training process, we only adopt the first two loss items in Eq. 13. As illustrated in Fig. 19, the results of **w/o pose** are temporally unstable. We observe several sudden changes in geometry, leading to the spikes of the curves in Fig. 19. Such an unstable phenomenon compromises the performance of facial animation, causing manual correction in post-production. Differently, our pose enhancement handles the head pose variance and alleviates the unstable artifact.

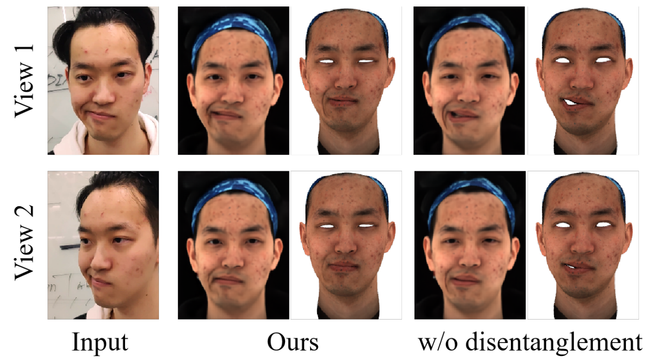



Fig. 18. **Qualitative evaluation of our expression disentanglement.** After training the network with the same dataset, we tested it with images from novel viewpoints. The variation without expression disentanglement suffers from inconsistency when the viewpoint changes. The absence of expression and viewpoint disentanglement leads to a reduced ability of generalization for our model. As a comparison, the results given by our model with expression disentanglement training handled viewpoint change with no deterioration in performance.

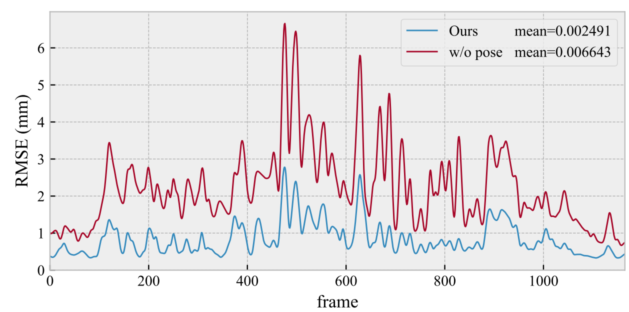



Fig. 19. **Quantitative evaluation of our pose augmentation.** We evaluate the stability of the predicted geometries with and without pose augmentations. The curve shows the mean distances of vertices to their stable coordinates, which are expected to be fixed. Our training scheme with pose augmentation leads to more stable face geometries, while the results given by the variation without it jitters while testing.

*Texture prediction strategy.* We further evaluate our wrinkle-map-based texture prediction strategy. Let **w/o wrinkle** denote our variation that directly predicts pixel-aligned textures instead of wrinkle maps and subsequently adopts an extra super-resolution network to upscale the predicted textures to 4K resolution. As shown in Fig. 20, our scheme models the wrinkle maps including shading maps, which enables better preservation of facial details. Our wrinkle map models the discrepancies on top of the neutral textures and can be scaled to high-resolution without suffering from blur or block-wise artifacts.

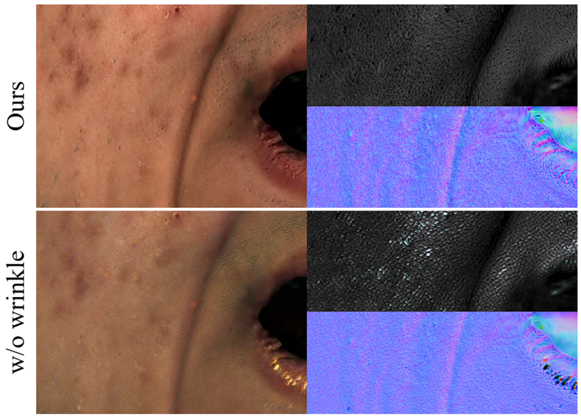



Fig. 20. **Qualitative evaluation of our texture predicting strategy.** Our variation without predicting wrinkles suffers from repeated artifacts and loses a lot of real details. The wrinkle map preserves more facial details and sharpens the textures and avoids scale-like artifact.

### 6.4 Limitations and Discussions

Though our neural physically-based facial asset has demonstrated its ability in a variety of applications with high-fidelity rendering performance, it is still subject to several limitations. Our approach lacks semantic animation controls. We choose implicit representation to model the expressions over explicit parametric methods, therefore the re-targeting results of the asset cannot be easily edited by an artist. While this remains a limitation to our method, there are alternatives: one can either replace our in-the-wild data with rendered images from parametric models analogous to [Moser et al. 2021] or directly map extracted explicit parameters from the in-the-wild images to our latent variables. Our approach achieves cross-identity facial motion re-targeting with robust expression disentanglement. Yet we capture the facial asset in a performer-specific paradigm, hence it can not well handle extremely challenging identity and environmental variance, similar to other performer-specific work [Laine et al. 2017; Lombardi et al. 2018]. Besides, our cross-identity facial motion retargeting relies on actor-specific training when applying to a new performer. It's an interesting direction to further design a generalizable expression network structure to simultaneously handle various identities, similar to those general facial models [Feng et al. 2021]. Moreover, we don't simulate the lighting conditions and background as well as previous methods (i.e. Klaudiny et al. [2017]). In future works, we plan to apply single-image relighting techniques with portrait matting (i.e. Zhang et al. [2021]) to simulate novel lighting environments.

Our geometry network only models soft tissues of the face with high fidelity, we still need other face components like teeth, gums, eyeballs, eye occlusion, and eyelashes, to improve the realism of our facial assets. Previous works apply different types of deformations, such as linear and laplacian deformations, to model the movements of these components, which may not be able to achieve high realism. In future works, we will incorporate eye movement capture and jaw adaptation algorithms into our model, and use a biologically sound parametric skull model as prior. Apart from face components, we also need realistic hair with correct deformations to build a production-level digital avatar. Rendering hair with high fidelity remains a challenging problem in computer graphics due to its complex physical forms. Existing solutions are mostly computationally expensive, yet recent work using neural radiance field [Mildenhall et al. 2020] to render fuzzy objects has shown promising results. Hence, we plan to further improve our method by rendering hair simultaneously using the method of [Luo et al. 2021].

In our practice, we separate textures into neutral textures and multi-channel wrinkle maps for modifying high-resolution textures. They can not simulate certain phenomena, such as pore squeezing. A possible solution is to apply a similar strategy as we use at the data processing stage, where we disentangle textures into warp fields and neutral textures to eliminate pixel drifts for alignment. We have tried such approach for modeling skin stretch and compression when generating textures, but failed due to jittering noise in the data. In future works, we will further improve our data processing pipeline, which upgrades the current wrinkle-map-based texture inference framework to a warp-field-based one. In the new framework, the dynamic texture generated by our network will be able to simulate pore-level deformation effects.

## 7 CONCLUSION

We have presented a neural approach to generate video-driven dynamic facial geometries with high-quality physically-based texture assets from FaStage data. Our neural approach enables high-fidelity performer-specific facial capture and even cross-identity facial motion re-targeting without manual labor. It also enables various physically-based facial editing effects, like geometry and material editing or wrinkle transfer. The core of our approach is a new multi-VAE-based design to disentangle facial attributes explicitly and effectively with the aid of a global MLP-based mapping across latent spaces. Our explicit expression and viewpoint disentanglement as well as the novel triplet semi-supervision scheme enable effective latent space construction for both the expression and geometry VAEs. Our wrinkle maps and view-independent texture inference in our physically-based texture VAE achieve high-quality 4K rendering of dynamic textures and convenient neutral-based editing for post-production. We showcase extensive experimental results and video-driven applications for high-fidelity facial animation, facial motion retargeting, and physically-based editing with high realism. We believe that our approach significantly improves the video-driven production-level workflow of generating dynamic facial geometries with physically-based assets. It serves as a critical step for automatic, high-quality, and controllable dynamic facial asset generation in the neural era, with many potential applications for believable digital human, film, game production, and immersive experience in VR/AR and the Metaverse.

## ACKNOWLEDGMENTS

This work was supported by Shanghai YangFan Program (21YF1429500), Shanghai Local College Capacity Building Program (22010502800), NSFC programs (61976138, 61977047), the National Key Research and Development Program (2018YFB2100500), STCSM (2015F0203-000-06) and SHMEC (2019-01-07-00-01-E00003).

## REFERENCES


Kfir Aberman, Rundi Wu, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. 2019. Learning Character-Agnostic Motion for Motion Retargeting in 2D. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 75.

Victoria Fernández Abrevaya, Adnane Boukhayma, Philip HS Torr, and Edmond Boyer. 2020. Cross-modal deep face normals with deactivable skip connections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4979–4989.

Linchao Bao, Xiangkai Lin, Yajing Chen, Haoxian Zhang, Sheng Wang, Xuefei Zhe, Di Kang, Haozhi Huang, Xinwei Jiang, Jue Wang, et al. 2021. High-Fidelity 3D Digital Human Head Creation from RGB-D Selfies. *ACM Transactions on Graphics (TOG)* 41, 1 (2021), 1–21.

Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. 2010. High-quality single-shot capture of facial geometry. In *ACM SIGGRAPH 2010 papers*. 1–9.

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 Transactions on Graphics (TOG)* 40, 4 (2021), 1–15.

Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 187–194.

Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. 2020. Neural head reenactment with latent pose descriptors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13786–13795.

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 Transactions on Graphics (TOG)* 40, 4 (2021), 1–17.

Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. 2015. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics (ToG)* 34, 4 (2015), 1–9.

Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, et al. 2022. Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–19.

Chen Cao, Yanlin Weng, Stephen Lin, and Kun Zhou. 2013. 3D shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG)* 32, 4 (2013), 1–10.

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2014. FaceWarehouse: A 3D Facial Expression Database for Visual Computing. *IEEE Transactions on Visualization and Computer Graphics* 20, 3 (2014), 413–425. https://doi.org/10.1109/TVCG.2013.249

Prashanth Chandran, Derek Bradley, Markus Gross, and Thabo Beeler. 2020. Semantic deep face models. In *2020 International Conference on 3D Vision (3DV)*. IEEE, 345–354.

Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. 2019. Photo-realistic facial details synthesis from single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9429–9439.

Lele Chen, Chen Cao, Fernando De la Torre, Jason Saragih, Chenliang Xu, and Yaser Sheikh. 2021. High-fidelity Face Tracking for AR/VR via Deep Lighting Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13059–13069.

Shiyang Cheng, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. 2018. 4DFAB: A Large Scale 4D Database for Facial Expression Analysis and Biometric Applications. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 5117–5126. https://doi.org/10.1109/CVPR.2018.00537

Darren Cosker, Eva Krumhuber, and Adrian Hilton. 2011. A FACS valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling. In *Proceedings of the IEEE International Conference on Computer Vision*. 2296–2303. https://doi.org/10.1109/ICCV.2011.6126510

Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 145–156.

Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 2020. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)* 39, 5 (2020), 1–38.

Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).

Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. 2021. Learning an Animatable Detailed 3D Face Model from In-The-Wild Images. *ACM Transactions on Graphics, (Proc. SIGGRAPH)* 40, 8. https://doi.org/10.1145/3450626.3459936

Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. 2018. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 534–551.

Graham Fyffe and Paul Debevec. 2015. Single-shot reflectance measurement from polarized color gradient illumination. In *2015 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 1–10.

Graham Fyffe, Tim Hawkins, Chris Watts, Wan-Chun Ma, and Paul Debevec. 2011. Comprehensive facial performance capture. In *Computer Graphics Forum*, Vol. 30. Wiley Online Library, 425–434.

Graham Fyffe, Andrew Jones, Oleg Alexander, Ryosuke Ichikari, and Paul Debevec. 2014. *Driving high-resolution facial scans with video performance capture.* Technical Report. UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES.

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*. 8649–8658.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised Domain Adaptation by Backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 37)*, Francis Bach and David Blei (Eds.). PMLR, Lille, France, 1180–1189. https://proceedings.mlr.press/v37/ganin15.html

Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. 2019. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1155–1164.

Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. 2011. Multiview face capture using polarized spherical gradient illumination. In *Proceedings of the 2011 SIGGRAPH Asia Conference*. 1–10.

Paulo Gotardo, Jérémy Riviere, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. 2018. Practical dynamic facial appearance modeling and acquisition. (2018).

Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. 2020. Towards Fast, Accurate and Stable 3D Dense Face Alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. 2019. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–19.

Yudong Guo, Jianfei Cai, Boyi Jiang, Jianmin Zheng, et al. 2018. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE transactions on pattern analysis and machine intelligence* 41, 6 (2018), 1294–1307.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2016. beta-vae: Learning basic visual concepts with a constrained variational framework. (2016).

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*. 172–189.

Loc Huynh, Weikai Chen, Shunsuke Saito, Jun Xing, Koki Nagano, Andrew Jones, Paul Debevec, and Hao Li. 2018. Mesoscopic facial geometry inference using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8407–8416.

Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. 2017. Panoptic Studio: A Massively Multiview System for Social Interaction Capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

Davis E King. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research* 10 (2009), 1755–1758.

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Martin Klaudiny, Steven McDonagh, Derek Bradley, Thabo Beeler, and Kenny Mitchell. 2017. Real-Time Multi-View Facial Capture with Synthetic Training. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 325–336.

Samuli Laine, Tero Karras, Timo Aila, Antti Herva, Shunsuke Saito, Ronald Yu, Hao Li, and Jaakko Lehtinen. 2017. Production-level facial performance capture using deep convolutional neural networks. In *Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation*. 1–10.

Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos P Zafeiriou. 2021. AvatarMe++: Facial Shape and BRDF Inference with Photorealistic Rendering-Aware GANs. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 01 (2021), 1–1.

Chloe LeGendre, Kalle Bladin, Bipin Kishore, Xinglei Ren, Xueming Yu, and Paul Debevec. 2018. Efficient multispectral facial capture with monochrome cameras. In *Color and Imaging Conference*, Vol. 2018. Society for Imaging Science and Technology, 187–202.

John P Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Frederic H Pighin, and Zhigang Deng. 2014. Practice and Theory of Blendshape Facial Models. *Eurographics (State of the Art Reports)* 1, 8 (2014), 2.

Jiaman Li, Zhengfei Kuang, Yajie Zhao, Mingming He, Karl Bladin, and Hao Li. 2020b. Dynamic facial asset and rig generation from a single scan. *ACM Trans. Graph.* 39, 6 (2020), 215–1.

Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Prasad, Bipin Kishore, Jun Xing, et al. 2020a. Learning formation of physically-based face attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3410–3419.

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 Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (2017), 194:1–194:17. https://doi.org/10.1145/3130800.3130813

Tianye Li, Shichen Liu, Timo Bolkart, Jiayi Liu, Hao Li, and Yajie Zhao. 2021. Topologically Consistent Multi-View Face Inference Using Volumetric Sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3824–3834.

Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep appearance models for face rendering. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–13.

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. https://doi.org/10.1145/3450626.3459863

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

H. Luo, A. Chen, Q. Zhang, B. Pang, M. Wu, L. Xu, and J. Yu. 2021. Convolutional Neural Opacity Radiance Fields. In *2021 IEEE International Conference on Computational Photography (ICCP)*. IEEE Computer Society, Los Alamitos, CA, USA, 1–12. https://doi.org/10.1109/ICCP51581.2021.9466273

Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. 2021. Pixel Codec Avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 64–73.

Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, Paul E Debevec, et al. 2007. Rapid Acquisition of Specular and Diffuse Normal Maps from Polarized Spherical Gradient Illumination. *Rendering Techniques* 2007, 9 (2007), 10.

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 *European conference on computer vision*. Springer, 405–421.

milesial. 2022. Pytorch-UNet. https://github.com/milesial/Pytorch-UNet.

Araceli Morales, Gemma Piella, and Federico M Sukno. 2021. Survey on 3D face reconstruction from uncalibrated images. *Computer Science Review* 40 (2021), 100400.

Lucio Moser, Chinyu Chien, Mark Williams, Jose Serra, Darren Hendler, and Doug Roble. 2021. Semi-supervised video-driven facial animation transfer for production. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–18.

Yuval Nirkin, Yosi Keller, and Tal Hassner. 2019. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*. 7184–7193.

Kyle Olszewski, Joseph J Lim, Shunsuke Saito, and Hao Li. 2016. High-fidelity facial and speech animation for VR HMDs. *ACM Transactions on Graphics (TOG)* 35, 6 (2016), 1–14.

Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. 2020. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems* 33 (2020), 7198–7211.

Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. 2009. A 3D face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*. Ieee, 296–301.

Amit Raj, Michael Zollhofer, Tomas Simon, Jason Saragih, Shunsuke Saito, James Hays, and Stephen Lombardi. 2021. Pixel-Aligned Volumetric Avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11733–11742.

Elad Richardson, Matan Sela, and Ron Kimmel. 2016. 3D face reconstruction by learning from synthetic data. In *2016 fourth international conference on 3D vision (3DV)*. IEEE, 460–469.

Jérémy Riviere, Paulo Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. 2020. Single-shot high-quality facial geometry and skin appearance capture. (2020).

Shunsuke Saito, Lingyu Wei, Liwen Hu, Koki Nagano, and Hao Li. 2017. Photorealistic facial texture inference using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5144–5153.

William AP Smith, Alassane Seck, Hannah Dee, Bernard Tiddeman, Joshua B Tenenbaum, and Bernhard Egger. 2020. A morphable face albedo model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5011–5020.

Giota Stratou, Abhijeet Ghosh, Paul Debevec, and Louis-Philippe Morency. 2011. Effect of illumination on automatic expression recognition: A novel 3D relightable facial database. In *2011 IEEE International Conference on Automatic Face Gesture Recognition (FG)*. 611–618. https://doi.org/10.1109/FG.2011.5771467

Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. 2017. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5163–5172.

Unity Technologies. 2020. The Heretic: Digital Human. https://assetstore.unity.com/packages/essentials/tutorial-projects/the-heretic-digital-human-168620.

Shih-En Wei, Jason Saragih, Tomas Simon, Adam W. Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. 2019. VR Facial Animation via Multiview Image Translation. 38, 4, Article 67 (jul 2019), 16 pages. https://doi.org/10.1145/3306346.3323030

Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. 2011. Realtime performance-based facial animation. *ACM transactions on graphics (TOG)* 30, 4 (2011), 1–10.

Chenglei Wu, Derek Bradley, Markus Gross, and Thabo Beeler. 2016. An anatomically-constrained local deformation model for monocular face capture. *ACM transactions on graphics (TOG)* 35, 4 (2016), 1–12.

Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. 2018. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–14.

Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. 2020. FaceScape: a Large-scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Baocai Yin, Yanfeng Sun, Chengzhang Wang, and Yun Ge. 2009. BJUT-3D large scale 3D face database and information processing. *Journal of Computer Research and Development* 46, 6 (2009), 1009.

Lijun Yin, Xiaochen Chen, Yi Sun, Tony Worm, and Michael Reale. 2008. A high-resolution 3D dynamic facial expression database. In *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*. 1–6. https://doi.org/10.1109/AFGR.2008.4813324

Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. 2006. A 3D facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*. IEEE, 211–216.

Jae Shin Yoon, Takaaki Shiratori, Shoou-I Yu, and Hyun Soo Park. 2019. Self-supervised adaptation of high-fidelity face models for monocular performance tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4601–4609.

Longwen Zhang, Qixuan Zhang, Minye Wu, Jingyi Yu, and Lan Xu. 2021. Neural Video Portrait Relighting in Real-Time via Consistency Modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 802–812.

Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. 2014b. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing* 32, 10 (2014), 692–706.

Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2014a. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*. Springer, 94–108.

Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. 2018. State of the art on monocular 3D face reconstruction, tracking, and applications. In *Computer Graphics Forum*, Vol. 37. Wiley Online Library, 523–550.