VariTex: Variational Neural Face Textures

Marcel C. Bühler[1] Abhimitra Meka[2] Gengyan Li[1,2] Thabo Beeler[2] Otmar Hilliges[1]
[1]ETH Zurich [2]Google
https://mcbuehler.github.io/VariTex

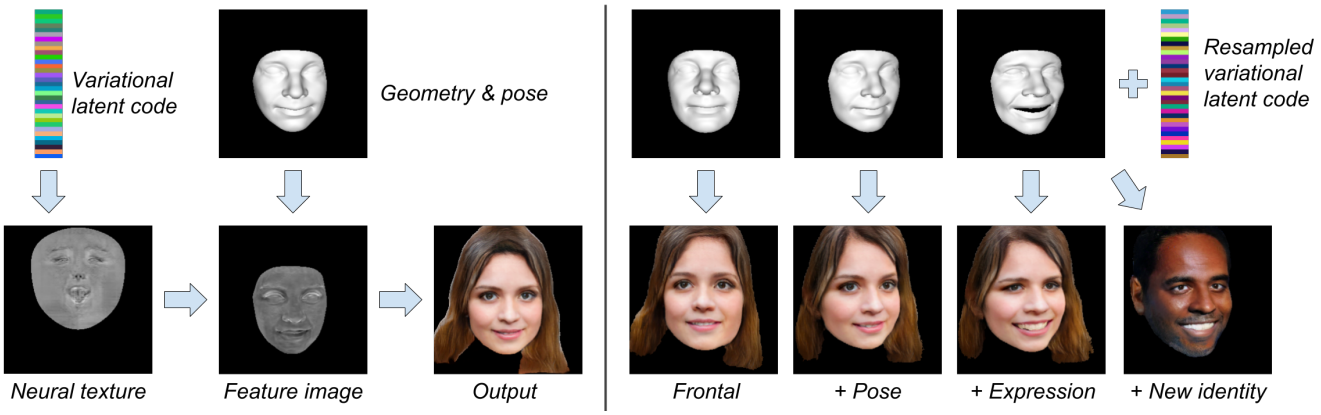

Figure 1. VariTex generalizes person-specific neural textures to variational textures. This allows to control both pose and expressions via explicit 3D geometries (Fig. 3) and to sample novel identities (Fig. 5). Our method generates images under fine head pose and expression control, while maintaining geometric consistency over a large range of these parameters (Fig. 4 and Tbl. 2).

Abstract

Deep generative models can synthesize photorealistic images of human faces with novel identities. However, a key challenge to the wide applicability of such techniques is to provide independent control over semantically meaningful parameters: appearance, head pose, face shape, and facial expressions. In this paper, we propose VariTex - to the best of our knowledge the first method that learns a variational latent feature space of neural face textures, which allows sampling of novel identities. We combine this generative model with a parametric face model and gain explicit control over head pose and facial expressions. To generate complete images of human heads, we propose an additive decoder that adds plausible details such as hair. A novel training scheme enforces a pose-independent latent space and in consequence, allows learning a one-to-many mapping between latent codes and pose-conditioned exterior regions. The resulting method can generate geometrically consistent images of novel identities under fine-grained control over head pose, face shape, and facial expressions. This facilitates a broad range of downstream tasks, like sampling novel identities, changing the head pose, expression transfer, and more.

1. Introduction

The ability to generate images with user-controlled parameters, such as identity-specific appearance, pose, and expressions would have many applications in computer graphics and vision. Synthesizing photorealistic images of novel human faces has recently been made possible through deep generative adversarial networks [17, 23, 24] or variational auto encoders [26], that learn the distribution of real faces to generate new identities. However, such methods typically do not provide semantic control over shape, pose, and facial expressions. This results in undesired global appearance changes across different generated images, for example, a change in identity when viewing from a different angle.

In order to gain more control over the generated images, recent work conditions neural networks on explicit 3D geometries [12, 16, 25, 27, 41, 42, 43]. Promising results have been shown by first generating the 2D face image from a learned latent space, and then attempting to rig it using graphics techniques in a geometrically consistent manner [41, 42]. This approach suffers from an inherent disadvantage: since the image synthesis is performed in 2D space only, it is hard to enforce consistency under 3D manipulation. Strong supervision via multi-view images or multi-pose data from monocular videos at *training* time can alleviate this to some degree. However, the inherent 2D na-

ture of the solution prohibits a truly 3D consistent solution for novel *test* poses, views, and expressions. This problem particularly manifests itself when synthesizing poses and expressions that lie outside the distribution of the training data (Fig. 4). To increase geometric faithfulness, recent work has attempted to learn the distribution of faces in 3D, for example by leveraging neural textures to represent 3D scenes in texture space [32, 43, 44]. Especially when trained from video, rendering 3D geometry with *neural* textures has been shown to produce highly consistent outputs for multiple poses and expressions, albeit at the cost of having to learn a texture *per subject*.

Our work generalizes subject-specific neural textures [18, 43, 44] to *variational* neural textures, enabling geometry-aware synthesis of *novel* identities (see Fig. 1 and 5). Neural textures represent the appearance of a 3D surface as 2D feature maps. In contrast to prior works, which are trained per subject, *variational* neural textures do not require strong supervision in the form of multi-view images or minutes-long video sequences as input *for each identity*. Instead, they are generated by sampling from an underlying latent distribution of neural face textures. Importantly, this latent space is learned in a *self-supervised* scheme from *monocular* RGB images *without requiring any annotations*. We use a parametric face model [1] in combination with a differentiable renderer to provide fine-grained control over face shape, head pose, and facial expression. This is sufficient to generate face interiors that preserve a subject's identity across pose and expression, but does not model other important details, such as ears, hair, and the mouth interior. To attain complete images of human heads, we propose a pose-aware additive decoder that generates features for visually plausible details (e.g., facial hair). We devise a novel training regime that allows the additive decoder to learn a one-to-many mapping and in consequence to generate the exterior face region conditioned on different head poses from the same latent code (see Fig. 1).

We propose VariTex: Variational Neural Face Textures – a method to sample novel identities and synthesize consistent faces in multiple poses and expressions (Fig. 1 and 3).

We demonstrate state-of-the-art (SoA) photo-realistic results for geometric control (Fig. 3), novel identity image synthesis (Fig. 5), and novel pose synthesis (Fig. 1 and 4). Our method achieves higher visual identity-consistency than related work (Fig. 4). Quantitatively, we compare embedding distances between frontal and posed faces via a SoA face recognition network [11] (Tbl. 2). Finally, we conduct a user study (Sec. 5.3), where participants rate consistency for posed faces and overall photo-realism.

In summary, we make the following contributions:

1. VariTex, the first method for learning a variational latent feature space for neural face textures - allowing to sample novel identities.
2. Combining the generative power of learned facial textures with the explicit control of a parametric face model enables fine-grained control over facial expressions, head pose, face shape, and appearance.
3. We synthesize plausible outputs for difficult regions where no 3D geometries are available (e.g., hair, ears, and the mouth interior).
4. We show that our method is more identity consistent under geometric transformations.

2. Related Work

We briefly review related work on image synthesis of human faces, particularly those that leverage differentiable rendering and neural textures.

Method	Pose-independent texture	Sampling novel identities
UV-GAN [9]	RGB	×
DNR [44]	neural	×
NVP [43]	neural	×
ConfigNet [27]	×	✓
GIF [16]	×	✓
DiscoFaceGAN [12]	×	✓
Ours	neural	✓

Table 1. Overview of most closely related methods. Texture-rendering based methods are not designed to sample new identities [9, 43, 44]. More generic synthesis methods [12, 16, 27] suffer from inconsistency under large pose variations (Fig. 4 and Tbl. 2) because they do not provide texture-level control over the face region. We propose a framework based on variational neural textures that can do both.

High-quality Face Synthesis. Most modern methods for synthesizing natural images leverage generative adversarial networks (GAN) [17] or variational auto-encoders [26]. These methods have achieved a high level of photorealism [2, 7, 8, 23, 24, 31, 34, 46]. Typically such methods learn to map from a low dimensional latent space to the distribution of 2D face images using convolutional neural networks. However, these latent spaces often entangle appearance and geometry [23, 24], making novel pose or expression synthesis extremely difficult. Recent works started disentangling the latent space and adding more and more control [4, 12, 14, 16, 22, 27, 39, 41, 42], for example, by learning disentanglement via statistical face models [1] as strong priors [12, 16]. Neural radiance fields [30] have shown to render faces of very high quality [13, 19, 33]. However, their generative variants [3, 36] still lack control over expressions. In summary, generative modeling of photorealistic faces with artistic control remains a difficult challenge.

Differentiable Rendering. An alternative way to disentangle appearance from geometry and pose *by design* is learning appearance in a UV space [9, 18, 29, 32, 38, 43, 44].

Traditional computer graphics rendering pipelines require highly detailed 3D geometries, which are very expensive to obtain. Recently, Thies et al. proposed *deferred neural rendering* [44]. Deferred neural rendering showed how deep neural networks can compensate imperfect 3D geometries and render highly photo-realistic imagery. Key components of these methods are *neural textures*. Instead of using traditional textures in a pre-defined *color space*, they leveraged the power of *neural features* as a description of texture. As a difference to related works based on neural textures [32, 44], our model is fully generative and allows sampling of new identities. Thies et al. [44] train person-specific models and Raj et al. [32] optimize person-specific textures from videos. Our model is trained from monocular images alone.

Neural Textures for Faces. Previous methods using neural textures [43, 44] learn a *person-specific* texture from multi-view images or videos. Given enough data of a target person, they enable realistic animation of the facial expressions seen during training. The high expression fidelity and image quality comes at the cost of tightly coupling neural textures and rendering, which requires training a network per person. Furthermore, training neural textures per scene requires multiple views or minute-long sequences of the target person. An interesting challenge in the evolution of neural textures is to generalize them to single images and to novel identities. To this end, we frame the problem as a *variational neural texture generation* task, followed by a texture-to-image translation task. This generalizes the texture and image generator to unseen identities, gives fine-grained control over head pose and facial expression and the generated images remain consistent under the manipulation of these parameters.

3. VariTex: Variational Neural Textures

3.1. Overview

We tackle the problem of controlled novel identity synthesis for faces, with the goal to disentangle appearance from pose and expression. To do so, we generalize person-specific neural textures [43, 44] to *variational* neural textures by learning a distribution over identities that can map to a neural texture space. This allows generating an infinite number of neural textures that can be mapped on face geometries with arbitrary poses and expressions.

At the core of our method is a neural texture decoder that is trained in a self-supervised manner via neural rendering. The decoder learns to generate neural textures that follow a predefined layout given by the UV parameterization of a 3D morphable face model [15], followed by a projection into image space and rendering as an RGB image.

Intuitively, our network can render extreme poses despite being trained on largely frontal imagery, because the neural texture projection provides spatially aligned features. Such neural rendering networks have been shown to generalize to poses unseen during training [44].

3.2. Problem Statement

Our goal is to learn a generator G_θ that produces face images $\hat{\boldsymbol{I}}$ and foreground masks $\hat{\boldsymbol{M}}$ from a latent description for identity $\boldsymbol{z} \in \mathbb{R}^{d_z}$ and control signals for shape $\boldsymbol{\alpha} \in \mathbb{R}^{d_\alpha}$, expression $\boldsymbol{\beta} \in \mathbb{R}^{d_\beta}$ and head pose $\boldsymbol{R} \in SO(3)$. Given a code for an identity $\boldsymbol{z}$ and a corresponding shape $\boldsymbol{\alpha}$, the generator should synthesize consistent images that preserve facial identity across different expressions $\boldsymbol{\beta}$ and poses $\boldsymbol{R}$. Equation 1 summarizes our problem statement:

$$(\hat{\boldsymbol{I}}, \hat{\boldsymbol{M}}) = G_\theta(\boldsymbol{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{R}). \tag{1}$$

The distribution over $\boldsymbol{z}$ is learned from a large collection of monocular face images. Shape $\boldsymbol{\alpha}$ and expression $\boldsymbol{\beta}$ are the coefficients of a PCA-based 3D morphable face model [1], learned from a collection of 3D scans [15]. The pose $\boldsymbol{R}$ is a 3D rotation matrix.

While we also train an image-to-latent-space encoder, we emphasize that this is more of a side effect. Our primary goal is to learn latent space from which *novel* identities can be sampled and rendered under geometric control, as opposed to generating novel views of existing identities.

3.3. Architecture Overview

Fig. 2 summarizes our method. During training, we use monocular RGB images to learn the underlying space of face appearance. This is done in the variational auto-encoder (VAE) framework [26], where an encoder learns to map input face images to parameters of a normal distribution. These parameters can then be sampled to generate a latent code which is interpreted by the VariTex generator $\boldsymbol{G}_\theta$. We describe our training scheme in Sec. 3.5.

Unlike a traditional decoder of a VAE [26], The VariTex Generator synthesizes face images in a geometry-aware manner. We use a parametric face model with consistent topology to map the 3D geometry of any face to a 2D texture layout. This 2D texture space serves as the domain over which feature maps for novel identities can be generated. The face model is then used to re-project the generated *neural textures* from this layout to the output image space under any desired pose and expression. We describe this process in greater detail in Sec. 3.4.

The texture layout can only handle regions of the face geometry that are present in the face model. We use an additional network—the *additive decoder*—to generate features for the exterior regions, such as hair, ears, and the mouth interior.

Finally, a neural renderer converts the neural features into an RGB image and a plausible *foreground mask*. The full generation process is described in detail in Sec. 3.4.

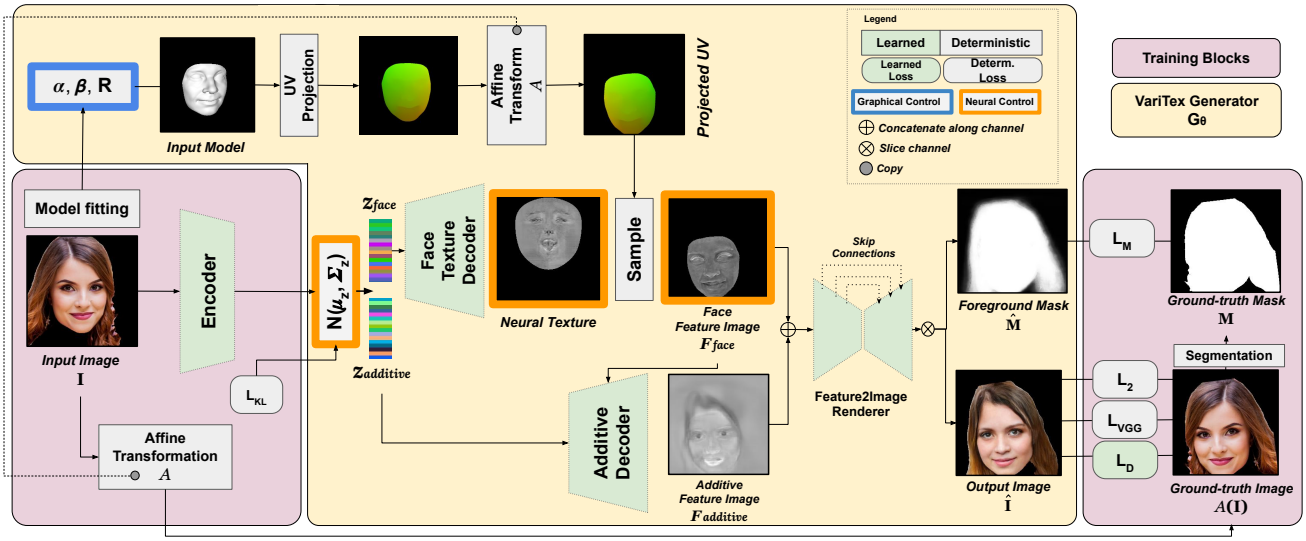

Figure 2. The objective of our pipeline is to learn a generator $\boldsymbol{G}_\theta$ that can synthesize face images with arbitrary novel identities whose expressions and pose can be controlled using face model parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$, and $\boldsymbol{R}$ (Fig. 3). During training, we use unlabeled monocular RGB images ($\boldsymbol{I}$) to learn a smooth latent space $\mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ of natural face appearance using a variational encoder. A latent code $\boldsymbol{z}$ sampled from this space is then decoded to a novel face image. At test time, we draw samples to generate novel face images (Fig. 5). Our variationally generated *neural textures* can also be stylistically interpolated to generate intermediate identities (supplementary material).

3.4. VariTex Generator

This section describes the components of the VariTex generator. The generator consists of two decoders and a Feature2Image rendering network. The decoders produce a neural description of the desired output—the *neural feature image*. The Feature2Image network turns these features into an RGB image and a corresponding foreground mask.

The generator allows a) to generate new identities by sampling latent codes $\boldsymbol{z}$ and shape coefficients $\boldsymbol{\alpha}$, and b) to manipulate expression $\boldsymbol{\beta}$ and pose $\boldsymbol{R}$.

The latent code for identity $\boldsymbol{z} \in \mathbb{R}^{256}$ can be sampled from a learned distribution $\mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ or extracted from a reference image. It is split into two halves: $\boldsymbol{z}_{face} \in \mathbb{R}^{128}$ for the face interior region, and $\boldsymbol{z}_{additive} \in \mathbb{R}^{128}$ for the regions outside the face model (e.g., hair). The latent code for the face $\boldsymbol{z}_{face}$ is converted by the face texture decoder into the face regions provided by the 3D model. The latent code for the rest of the head $\boldsymbol{z}_{additive}$ is processed into features for the rest of the face by the additive decoder.

The coefficients for shape $\boldsymbol{\alpha}$ and expression $\boldsymbol{\beta}$ can be sampled from a distribution extracted from reference images via 3D model fitting [15], or specified manually, which allows artistic control.

Face Texture Decoder. The face texture decoder is a modified ResNet-18 [20] where we expand the latent code $\boldsymbol{z}_{face}$ to spatial feature maps and stack them along the channel dimension. The feature maps are processed in a series of upsampling and residual blocks to the desired texture dimensions. The output is a pose and shape independent multi-dimensional feature map in UV space, which we call *neural texture*. We provide the detailed architecture in the supplementary material.

UV Rendering and Texture Sampling. In order to project the texture onto the image plane, we use a 3D morphable face model with a UV parameterization [15]. Given model coefficients for shape $\boldsymbol{\alpha}$, expression $\boldsymbol{\beta}$, and a rotation matrix $\boldsymbol{R}$, we compute the posed mesh. We then project the UV parameterization to image space following the standard computer graphics pipeline and use it to sample features from the neural face texture. The output of this step is a *neural face feature image* $\boldsymbol{F}_{face}$.

Additive Decoder. The face texture decoder yields a neural texture for the face region only. The additive decoder adds features for the regions missing in the face model, e.g., the hair or mouth interior. This is a very challenging task because the shape and appearance of the added regions should be consistent even for extreme head poses. The additive decoder should therefore be invariant to pose-dependent features in the latent code. Please refer to Sec. 3.5 and the supplementary material for more details.

We condition the additive decoder on both the latent description of identity $\boldsymbol{z}_{additive}$ and the neural face feature image $\boldsymbol{F}_{face}$. The latent code $\boldsymbol{z}_{additive}$ is expanded to a spatial feature map (similar to the face texture decoder) and upscaled in a series of ResNet layers [20]. In each block, we concatenate the rescaled face feature image as conditioning on geometry and pose.

The output of the additive texture decoder is an *additive*

feature image $\boldsymbol{F}_{additive}$ that is pixel-aligned given a pose, shape, expression, and identity.

Feature2Image Network. The last step of the VariTex Generator pipeline is to convert the feature images $\boldsymbol{F}_{face}$ and $\boldsymbol{F}_{additive}$ to an RGB output image. The Feature2Image network translates the stacked feature images into an RGB image and a foreground mask. Similar to [43, 44], the Feature2Image network is a U-Net [35].

3.5. Training

In contrast to existing methods that require strong supervision in the form of multi-view images or videos, we train only on unpaired monocular RGB images.

Encoder. During training, we learn a latent space $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$. A ResNet-18 [20] encoder takes a foreground-masked RGB image and predicts the mean $\boldsymbol{\mu}_z \in \mathbb{R}^{256}$ and diagonal covariance $\boldsymbol{\Sigma}_z \in \mathbb{R}^{256\times256}$, from which we sample a latent code $\boldsymbol{z} \in \mathbb{R}^{256}$ and process it further as described in Sec. 3.4.

Augmentation Scheme. While the parametric face model allows for geometry-consistent synthesis for the face interior region, doing the same for the face exterior, where no 3D geometry is available, is much more challenging. A Variational Auto Encoder [26] trained by reconstruction would simply learn to copy such regions (e.g., hair) into the same spatial location even under different poses.

To solve this problem, we employ an augmentation scheme to map our input image $\boldsymbol{I}$ to a transformed output image $A(\boldsymbol{I})$. The mapping A consists of random affine transforms: in-plane rotation, translation, scaling, and flipping. As a result, the additive decoder is guided to learn a one-to-many mapping—the same latent code $\boldsymbol{z}_{additive}$ must yield different additive feature images, which is determined by the pose and geometry from the face feature image. Please see Fig. 2 for a visual example and the supplementary for more details.

Objective Function. Each training sample consists of a foreground-masked training image $\boldsymbol{I}$, its affine transformed version $A(\boldsymbol{I})$, the ground-truth segmentation mask $\boldsymbol{M}$ belonging to $A(\boldsymbol{I})$, and their corresponding reconstructions $\hat{\boldsymbol{I}}$ and $\hat{\boldsymbol{M}}$. We denote the spatial dimensions as H and W.

For self-supervised reconstruction, we employ a photometric $\mathcal{L}_2$ loss term and a perceptual loss term $\mathcal{L}_{VGG}$:

$$\begin{aligned}\mathcal{L}_2 =& \|\hat{\boldsymbol{I}} - A(\boldsymbol{I})\|_2^2,\\ \mathcal{L}_{VGG} =& \sum_j v_j \|\phi_{VGG_j}(\hat{\boldsymbol{I}}) - \phi_{VGG_j}(A(\boldsymbol{I}))\|_1,\end{aligned} \tag{2}$$

where the function $\phi_{VGG_j}(\cdot)$ extracts the j-th feature map from a pretrained VGG network [10, 40], and v_j are the weights per feature map (listed in the supplementary).

In order to learn correct foreground masks, we supervise with a cross entropy loss term $\mathcal{L}_M$:

$$\begin{aligned}\mathcal{L}_M =& -\frac{1}{HW}\sum_i^H\sum_j^W \boldsymbol{M}_{ij}\log\hat{\boldsymbol{M}}_{ij}\\ &+ (1-\boldsymbol{M}_{ij})(1-\log\hat{\boldsymbol{M}}_{ij}).\end{aligned} \tag{3}$$

We smooth the latent space with a Kullback-Leibler regularization term:

$$\mathcal{L}_{KL} = \mathcal{D}_{KL}\left(q(\boldsymbol{z}|\boldsymbol{I})||p(\boldsymbol{z})\right), \tag{4}$$

where $q(\boldsymbol{z}|\boldsymbol{I})$ is the distribution predicted by the encoder and $p(\boldsymbol{z})$ is a standard Gaussian distribution [26].

To encourage realism, we employ a two-scale patch discriminator D [31] with feature matching. The adversarial generator loss term is

$$\begin{aligned}\mathcal{L}_{adv} =& (1-D(\hat{\boldsymbol{I}}))^2\\ &+ \sum_j ||\phi_{D_j}(\hat{\boldsymbol{I}}) - \phi_{D_j}(A(\boldsymbol{I}))||_1,\end{aligned} \tag{5}$$

where the function $\phi_{D_j}(\cdot)$ extracts the j-th feature map from the discriminator network.

The final losses for the generator and discriminator are:

$$\begin{aligned}\mathcal{L}_{\text{Generator}} =& \lambda_2\mathcal{L}_2 + \lambda_{VGG}\mathcal{L}_{VGG} + \lambda_M\mathcal{L}_M\\ &+ \lambda_{KL}\mathcal{L}_{KL} + \lambda_{adv}\mathcal{L}_{adv},\\ \mathcal{L}_{\text{Discriminator}} =& \lambda_{adv}\frac{1}{2}\left[D(\hat{\boldsymbol{I}})^2 + (1-D(A(\boldsymbol{I})))^2\right].\end{aligned} \tag{6}$$

We empirically choose $\lambda_2 = \lambda_M = \lambda_{adv} = 1$, $\lambda_{VGG} = 2$, and $\lambda_{KL} = 0.1$. For more training details and hyperparameters, please refer to the supplementary material.

4. Experimental Setup

Data and Preprocessing. We train our method on face images from the Flickr-Faces-HQ dataset (FFHQ) [23]. For training, we fit the Basel Face Model [15] offline. In nine cases, the model fitting fails ($< 0.02\%$ of all images). We remove those images from the training set and end up with $59,991$ training and $10,000$ test samples, following the recommended splits. We visualize the removed images in the supplementary material.

We aim to generate images with their corresponding foreground masks. To get pseudo-ground-truth, we train a state-of-the-art face segmentation network [5, 6] on CelebAMask-HQ [28] and predict the segmentation maps offline. Please refer to the supplementary material for details.

Identity Consistency Metric. To evaluate identity consistency, we compute a similarity score from the embeddings

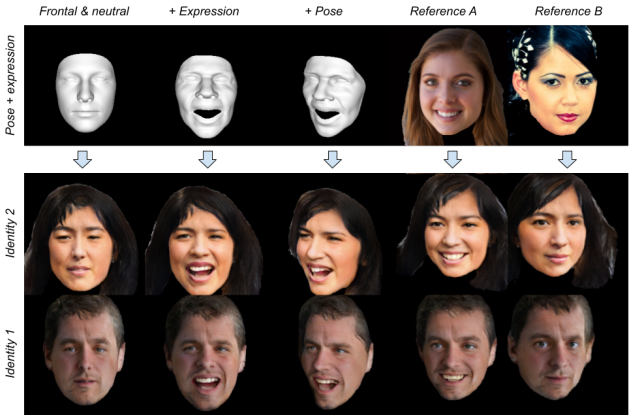

Figure 3. Rendering two identities under expression and pose control. Column 1 starts with a neutral pose and expression. Columns 2 and 3 change expression and pose via the graphical control unit (Fig. 2). For columns 4 and 5, we render the face with expression and pose from real reference images. The top row shows the corresponding face meshes and reference images.

of a state-of-the-art face recognition network [11]. For each related method, we render 3,000 identities with frontal head pose and compute their embeddings [11]. We then re-pose the same identities to various degrees and compute the cosine similarity between the normalized embeddings. As a reference for the reader, we provide similarities of a real-world multi-view dataset [45]. The real-world dataset contains faces with a slightly non-frontal pose (about $\pm 7°$), hence, we use the average embedding of the two most frontal faces for the frontal pose.

5. Results and Discussion

This section discusses our results for controlled face synthesis. In Sec. 5.1, we show qualitative results for rendering different geometries and poses (Fig. 3) and sampling novel identities (Fig. 5). In Section 5.2, we compare both qualitatively (Fig. 4) and quantitatively (Tbl. 2) with related work. In Sections 5.3 and 5.4, we conduct a user and an ablation study. In Section 5.5, we discuss limitations and future work.

5.1. Qualitative Results

Controlling Geometry and Pose. VariTex can sample novel identities and produce consistent images for different geometries and poses. Fig. 3 shows sequential edits on two identities. We start at a frontal pose and a neutral expression (column 1). We use the graphical control unit (Fig. 2) to change expression and pose (columns 2 and 3). The top row shows the corresponding face mesh. It is also possible to extract the parameters of the graphical control unit from reference images (columns 4 and 5). Our method maintains high identity consistency across manipulations.

Sampling and Identity Mixing. While previous works are limited to person-specific face textures [43, 44], VariTex can generate textures for novel identities by sampling in a latent space (Fig. 2). In Fig. 5, we sample new variants $\boldsymbol{z}_j \sim N(\boldsymbol{\mu}_{z_j}, \boldsymbol{\Sigma}_{z_j})$ for identities j from the test set. VariTex can also interpolate between two latent codes. We show such examples in the supplementary document and video.

5.2. Identity Consistency

A key benefit of using textured 3D geometries is that they allow highly consistent renderings even for extreme head poses. Our method leverages the strict mapping from texture to image space (Sec. 3.4). This facilitates the rendering of the identity-specific facial appearance for extreme poses, despite a dataset with mostly frontal faces. We visualize the head pose distribution of the training set and out-of-distribution samples in the supplementary.

We visually compare identity consistency in Fig. 4. Related works [12, 16, 27, 41] achieve highly consistent and photo-realistic results for frontal faces and poses up to $30°$ (pitch) and $15°$ (yaw). For more extreme poses, they tend to show severe artifacts [12, 16, 41] or blurred results [27].

For StyleRig [41], we exclusively show qualitative results because only a handful sample images were available to us. For the other methods [12, 16, 27] we generate 3000 samples and conduct a quantitative comparison by computing a similarity score using the identity consistency metric (Sec. 4). Tbl. 2 lists the resulting similarity scores (higher is better). Our method achieves the highest similarities, except for one of the evaluated poses.

5.3. User Study

We conduct a perceptual user-study comparing our method with three state-of-the-art techniques for controlled face image synthesis [12, 16, 27] along two dimensions:

1. The general quality of **photorealism** produced by the methods for images posed at random variations in the range of $[-45°, 45°]$ from the frontal pose. Participants answered the following question for 20 randomly chosen image pairs: *Of the two images, which looks more like a real person?*
2. **Identity consistency** for triplets of images of the same identity synthesized at 3 different poses: frontal pose, -45° and 45° degrees along the yaw and pitch axis. Each user was shown 10 randomly chosen pairs of such triplet images generated by ours and the related works. We asked: *Which set represents the same person more consistently?*

The survey consisted of 128 participants. Our method is on par in terms of photorealism, and clearly outperforms competing baselines for identity consistency criteria. In the following, we report the user study results for pairwise comparisons of each related work against ours.

Figure 4. Comparison with related work. GIF [16], ConfigNet [27], and DiscoFaceGAN (DFG) [12] achieve impressive visual quality for re-posing faces, but only up to 15° from the frontal pose. StyleRig [41] renders photorealistic outputs, but is unable to render strong pose variations and instead falls back to a smaller pose variation value, for example as seen in the $+45°$ case. Our technique is capable of synthesizing more *extreme* poses while maintaining high identity consistency with the frontal image.

Method	Similarity yaw ↔							Similarity pitch ↕						
	-45°	-30°	-15°	0°	15°	30°	45°	-45°	-30°	-15°	0°	15°	30°	45°
[27]	0.208	0.509	0.790	-	0.795	0.515	0.257	≤ 0	0.014	0.459	-	0.476	0.095	≤ 0
[16]	0.133	0.264	0.485	-	0.487	0.257	0.117	0.039	0.164	0.400	-	0.448	0.191	0.095
[12]	0.530	0.690	0.866	-	0.863	0.675	0.521	0.270	0.461	0.781	-	**0.826**	0.581	0.388
Ours	**0.568**	**0.729**	**0.874**	-	**0.873**	**0.732**	**0.585**	**0.416**	**0.611**	**0.821**	-	0.817	**0.611**	**0.420**
Ref [45]	0.855	0.845	0.726	-	0.790	0.773	0.779	0.719	0.725	0.753	-	0.797	0.805	0.782

Table 2. Identity consistency for different head poses. We compare 3,000 frontal faces (0°) with randomly sampled expressions with their respective posed variants. The scores indicate the similarity calculated as the dot product between normalized embeddings from a state-of-the-art face recognition network [11] (higher is better). The bottom row (*Ref*) is a reference to a real-world multi-view dataset [45]. For a visual comparison, please refer to Fig. 4.

For photorealism, 10% of the participants voted in favour of ConfigNet [27], 35% preferred GIF [16] and 50% chose DiscoFaceGAN [12]. For identity consistency, 0% of the participants preferred ConfigNet or GIF against VariTex; 8% of participants preferred DiscoFaceGAN. We provide results on other poses a random selection of example images from the survey in the supplementary document.

5.4. Ablation Study

We analyze the effect of neural textures in an ablation study. We simulate a traditional RGB texture by limiting the texture to 3 dimensions and imposing an additional con-

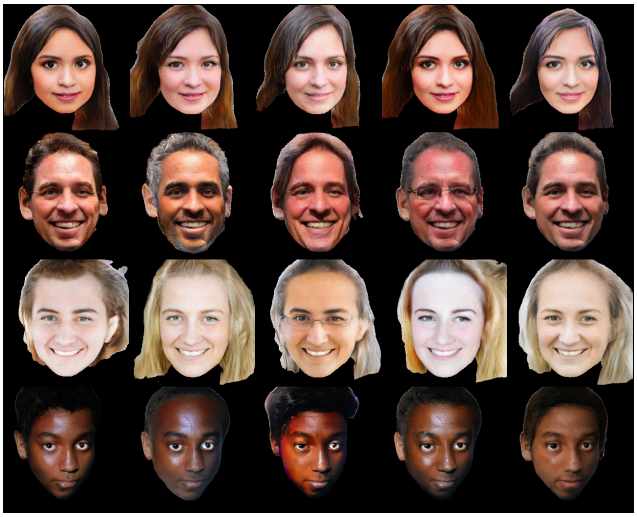

Figure 5. Sampling new identities. Each row samples from the learned latent distribution to generate variants of an identity. Note that the expression and pose are highly consistent.

Variant	FID↓	Consistency (yaw) ↑
3-dim texture with $\mathcal{L}_{RGB}$	54.27	0.712 ± 0.123
3-dim texture w/o $\mathcal{L}_{RGB}$	47.87	0.684 ± 0.132
16-dim with $\mathcal{L}_{RGB}$	37.96	0.724 ± 0.119
16-dim w/o $\mathcal{L}_{RGB}$ (**Ours**)	**34.35**	$\mathbf{0.727} \pm 0.121$

Table 3. Ablation study. We compare photorealism (FID [21, 37]) and identity consistency for *neural* vs. *RGB* textures. A higher dimensional neural texture can yield photorealistic outputs, while also maintaining high consistency. We provide detailed ablation results and visual examples in the supplementary material.

straint to make the texture resemble a classical RGB texture: $\mathcal{L}_{RGB} = \frac{1}{3}\sum_{c=1}^{3} ||\boldsymbol{F}_c - A(\boldsymbol{I}_c)||_2^2$.

The variable $\boldsymbol{F}$ denotes the *feature image* (Fig. 2) and $A(\boldsymbol{I})$ denotes the masked affine transformed training image (as described in Sec. 3.5). The subscript $c = 1, ..., 3$ represents the three RGB channels.

We train four combinations: a) a 3-dimensional texture with $\mathcal{L}_{RGB}$, b) a 3-dimensional *neural* texture without $\mathcal{L}_{RGB}$, c) a 16-dimensional texture with $\mathcal{L}_{RGB}$ and d) a 16-dimensional *neural* texture without $\mathcal{L}_{RGB}$ (**ours**).

Table 3 compares photorealism (FID [21, 37]) and identity consistency over the head poses (as described in Sec. 5.2). The consistency scores are the mean and the corresponding standard deviations over all poses. Please note that FID is computed on images masked to the foreground—the values are not directly comparable to related works that use backgrounds. Tree dimensional textures yield lower consistency and the generated images show artifacts—mostly visible in difficult regions, like eyes. The results indicate that our network benefits from the higher expressiveness of neural textures. The FID score shows that high-dimensional textures improve realism. In the supplementary material, we provide additional results and further ablations.

5.5. Limitations and Future Work

The proposed architecture allows going outside the training distribution. However, we observe a significant decrease in performance at very extreme poses beyond 60°. Furthermore, rigid objects inside the face get distorted by the perspective projection, e.g., when re-posing a face with glasses. We demonstrate examples for both cases in the supplementary material.

Possible extension to this work could generate complete images including backgrounds and torsos, and further disentangle the latent identity space.

6. Conclusion

We introduce VariTex—a generative model of neural face textures. The VariTex framework affords sampling novel identities while controlling both pose and geometry. Previous works excelled at either task individually; our framework generates novel identities *and* renders them in a significantly larger range of controlled poses and expressions. Our method achieves this by learning to synthesize an arbitrary pose-independent neural texture from a latent code, sampled from a distribution that is learned in a fully self-supervised scheme from monocular face images. The neural texture is then rendered to an image with any desired pose and expression. Our method also consistently generates the challenging face exterior regions such as hair, ears, and mouth-interiors. We demonstrate the capabilities of our method through qualitative, quantitative, and perceptual analysis. We also identify the limitations and discuss the various possibilities emerging from this line of work.

Acknowledgments and Disclosure of Funding. We thank Xucong Zhang, Emre Aksan, Thomas Langerak, Xu Chen, Mohamad Shahbazi, Velko Vechev, Yue Li, and Arvind Somasundaram for their contributions. We also thank Ayush Tewari for the StyleRig visuals. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program grant agreement No 717054.

References

[1] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 2, 3

[2] Marcel C Buhler, Andrés Romero, and Radu Timofte. Deepsee: Deep disentangled semantic explorative extreme super-resolution. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

[3] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5799–5809, 2021. 2

[4] Anpei Chen, Ruiyang Liu, Ling Xie, Zhang Chen, Hao Su, and Yu Jingyi. Sofgan: A portrait image generator with dynamic styling. *ACM Trans. Graph.*, 41(1), 2021. 2

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 5

[6] Liang Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Rethinking atrous convolution for semantic image segmentation liang-chieh. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 5

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 2

[8] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020. 2

[9] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7093–7102, 2018. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 2, 6, 7

[12] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *IEEE Computer Vision and Pattern Recognition*, 2020. 1, 2, 6, 7

[13] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 2

[14] Stephan J Garbin, Marek Kowalski, Matthew Johnson, and Jamie Shotton. High resolution zero-shot domain adaptation of synthetically rendered face images. In *European Conference on Computer Vision*, pages 220–236. Springer, 2020. 2

[15] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE, 2018. 3, 4, 5

[16] Partha Ghosh, Pravir Singh Gupta, Roy Uziel, Anurag Ranjan, Michael J. Black, and Timo Bolkart. GIF: Generative interpretable faces. In *International Conference on 3D Vision (3DV)*, 2020. 1, 2, 6, 7

[17] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 1, 2

[18] Artur Grigorev, Karim Iskakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. Stylepeople: A generative model of fullbody human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5151–5160, 2021. 2

[19] Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 5

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017. 8

[22] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Proc. NeurIPS*, 2020. 2

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1, 2, 5

[24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 1, 2

[25] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt.

Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 1

[26] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 1, 2, 3, 5

[27] Marek Kowalski, Stephan J. Garbin, Virginia Estellers, Tadas Baltrušaitis, Matthew Johnson, and Jamie Shotton. Config: Controllable neural face image generation. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 6, 7

[28] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020. 5

[29] Abhimitra Meka, Rohit Pandey, Christian Haene, Sergio Orts-Escolano, Peter Barnum, Philip Davidson, Daniel Erickson, Yinda Zhang, Jonathan Taylor, Sofien Bouaziz, Chloe Legendre, Wan-Chun Ma, Ryan Overbeck, Thabo Beeler, Paul Debevec, Shahram Izadi, Christian Theobalt, Christoph Rhemann, and Sean Fanello. Deep relightable textures - volumetric performance capture with neural rendering. *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia)*, 39(6), December 2020. 2

[30] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020. 2

[31] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 2, 5

[32] Amit Raj, Julian Tanke, James Hays, Minh Vo, Carsten Stoll, and Christoph Lassner. ANR: articulated neural rendering for virtual avatars. *CoRR*, abs/2012.12890, 2020. 2, 3

[33] Amit Raj, Michael Zollhofer, Tomas Simon, Jason Saragih, Shunsuke Saito, James Hays, and Stephen Lombardi. Pixel-aligned volumetric avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11733–11742, 2021. 2

[34] Andrés Romero, Pablo Arbeláez, Luc Van Gool, and Radu Timofte. Smit: Stochastic multi-label image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019. 2

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5

[36] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[37] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.1.1. 8

[38] Gil Shamai, Ron Slossberg, and Ron Kimmel. Synthesizing facial photometries and corresponding geometries using generative adversarial networks. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(3s):1–24, 2019. 2

[39] Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, and Gerard Medioni. Gan-control: Explicitly controllable gans. *arXiv preprint arXiv:2101.02477*, 2021. 2

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 5

[41] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images, cvpr 2020. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, june 2020. 1, 2, 6, 7

[42] Ayush Tewari, Mohamed Elgharib, Mallikarjun BR, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. Pie: Portrait image embedding for semantic control. *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia)*, 39(6), December 2020. 1, 2

[43] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *European Conference on Computer Vision*, pages 716–731. Springer, 2020. 1, 2, 3, 5, 6

[44] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2, 3, 5, 6

[45] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision*, pages 365–381. Springer, 2020. 6, 7

[46] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020. 2