UV-IDM: Identity-Conditioned Latent Diffusion Model for Face UV-Texture Generation

Hong Li[1]*, Yutang Feng[1]*, Song Xue[2], Xuhui Liu[1], Bohan Zeng[1]
Shanglin Li[1], Boyu Liu[1], Jianzhuang Liu[3], Shumin Han[2], Baochang Zhang[1,4,5†]
[1]Beihang University [2]Baidu VIS [3]Shenzhen Institute of Advanced Technology, Shenzhen, China
[4]Zhongguancun Laboratory, Beijing, China [5]Nanchang Institute of Technology, Nanchang, China

Abstract

*3D face reconstruction aims at generating high-fidelity 3D face shapes and textures from single-view or multi-view images. However, current prevailing facial texture generation methods generally suffer from low-quality texture, identity information loss, and inadequate handling of occlusions. To solve these problems, we introduce an **I**dentity-Conditioned Latent **D**iffusion **M**odel for face **UV**-texture generation (UV-IDM) to generate photo-realistic textures based on the Basel Face Model (BFM). UV-IDM leverages the powerful texture generation capacity of a latent diffusion model (LDM) to obtain detailed facial textures. To preserve the identity during the reconstruction procedure, we design an identity-conditioned module that can utilize any in-the-wild image as a robust condition for the LDM to guide texture generation. UV-IDM can be easily adapted to different BFM-based methods as a high-fidelity texture generator. Furthermore, in light of the limited accessibility of most existing UV-texture datasets, we build a large-scale and publicly available UV-texture dataset based on BFM, termed BFM-UV. Extensive experiments show that our UV-IDM can generate high-fidelity textures in 3D face reconstruction within seconds while maintaining image consistency, bringing new state-of-the-art performance in facial texture generation.*

1. Introduction

In recent years, the importance of facial digitization in the fields of VR, AR, and film has become increasingly prominent. The task of reconstructing the 3D shape and texture of a face from one in-the-wild image is a significant and difficult challenge within the fields of computer vision and graphics. In the past two decades, active research efforts have been devoted to the effective reconstruction through parametric fitting with a linear statistical model known as the 3D Morphable Model (3DMM), which was first introduced by Blanz and Vetter [5]. To enhance the quality of facial reconstruction, a central challenge is generating realistic facial textures from the original images. This can be primarily accomplished through linear and nonlinear methods. The former [9, 11, 13, 57] tends to generate low-quality reconstruction textures and is unable to capture the high-frequency features of in-the-wild images due to the limited expressive power of linear models. The latter attempts to utilize the nonlinear fitting ability of Deep Neural Networks (DNNs), such as Generative Adversarial Networks (GANs), to generate realistic textures. For example, [62, 63] use an encoder-decoder structure to generate position and texture maps, but they lose too much detailed information and have high convergence difficulty. [13, 54] attempt to achieve detailed textures indirectly by generating geometry structures with high-frequency details. GAN-based methods [4, 14, 16, 34–36, 73] commonly combine iterative optimization [10] to generate high-resolution textures. However, GANs may cause identity leakage due to over-smoothing, while iterative optimization approaches in practice are prone to incurring high time costs and overfitting from occlusion. Relightify [46] first demonstrates that image translation through an unconditional diffusion model can address the texture generation problem, thereby improving both realism and efficiency simultaneously.

In this paper, we introduce an Identity Conditional Latent Diffusion Model for facial UV texture generation (UV-IDM) as our texture generator to generate high-fidelity textures from individual in-the-wild images. In specific, we treat texture generation as a texture-completion task. We first extract incomplete textures from in-the-wild images using visibility masks and UV mapping relationships. Then, to fully utilize the features provided by the incomplete textures, we design an identity-conditioned module (ICM) to encode them and use the cross-attention mechanism to guide the generation process. This enhances the ability of UV-IDM to preserve identity and details. Our method

*These authors contributed equally.
†Corresponding Author: bczhang@buaa.edu.cn.

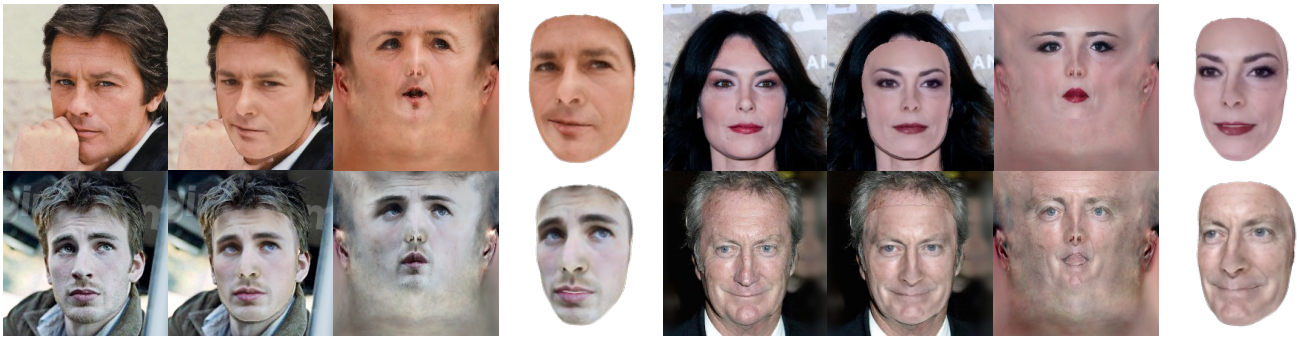

Figure 1. Examples of texture images produced by UV-IDM incorporated in Deep3D. For each group, from left to right: real face images; rendering result for generated texture; generated texture in UV space; and the color of the texture mapped to vertices in 3D space.

can be easily embedded as a plug-and-play module into face reconstruction methods based on the Basel Face Model (BFM) [47] to serve as a texture generator. It leverages the powerful generative capability of the Latent Diffusion Model (LDM) [50] to help achieve high-quality texture generation. As shown in Fig. 1, our texture generation method has a good generalization to occlusions (the fist in the top-left corner and the hair in the top-right corner), gender, and age in the input photos. The reconstructed texture performs well in preserving the identity of captured images and exhibits high fidelity and fine-grained details, such as the eye socket in the bottom-left corner and the highlight in the bottom-right corner.

[34–36, 46] obtain thousands of UV texture maps as training data by capturing from the real world using polarization spherical gradient illumination devices [17]. However, due to the high cost of such devices, it is difficult to implement large-scale data collection, resulting in a limited dataset size and certain limitations in the model's generalization ability. FFHQ-UV [4] synthesizes a high-quality publicly available UV texture dataset using HIFI3D++ [6] and StyleFlow [1], but the dataset cannot be reproduced due to copyright restrictions. Directly transferring this dataset to the more widely used BFM-based facial reconstruction methods [9, 38] still presents certain challenges.

Therefore, we would like to train an end-to-end generator based on the BFM to generate high-quality textures from in-the-wild images. This generator needs to recreate the identity faithfully and recover the illumination effectively while handling complex expressions and exhibiting robustness to hair occlusions. To train it, we use the normalized three-view data provided by FFHQ-UV [4] (eliminating illumination, hair, and expressions) to generate a set of high-quality texture data on the BFM. However, FFHQ-UV only provides normalized images, making it impossible to generate paired data from in-the-wild images to textures. Moreover, the process of converting in-the-wild images to standardized three-view images becomes non-replicable as the API provided by Microsoft is no longer publicly available. Therefore, we additionally leverage the editability [55, 68] of StyleGAN2 [30] to further expand the set by more than 80K high-fidelity "in-the-wild image, hair removal image, UV texture map" triples. These triples include various disruptive factors such as highlights, expressions, and hairs, thus meeting the training requirements for generating high-fidelity UV textures directly from a single in-the-wild image. The collection of data is called the BFM-UV dataset.

To summarize, our main contributions are as follows:

- We propose a facial texture generator based on an identity-guided latent diffusion model (UV-IDM), which can generate a high-fidelity texture UV image in just a few seconds.
- We supply a large-scale, high-fidelity BFM-based UV texture dataset encompassing more than 80K subjects and disclose the manufacturing process to enable researchers to use or generate textures based on other 3DMM.
- We conduct extensive experiments on three widely used datasets for 3D face reconstruction tasks. UV-IDM achieves state-of-the-art performance in both qualitative and quantitative results, generating high-quality and faithful facial textures from complex wild images.

2. Related Work

3D face reconstruction. 3D face reconstruction [9, 12, 19, 63, 75, 83] aims to restore 3D face shapes, expressions and textures from 2D face images. The 3D Face Morphable Model (3DMM) [5] represents a linear combination of shape and texture based 3D face statistical models. Under the 3DMM framework, current methods [11, 84] can be divided into either optimization-based fitting [25, 81] or learning-based regression [9, 19, 51, 60]. Optimization-based methods can yield precise results but also take expensive time. With the development of deep learning, neural networks have made new advances in estimating 3DMM parameters [9, 13, 19]. The original 3DMM model is limited to a low-dimensional linear space, resulting in a lack of reconstruction detail. Learning nonlinear substrates or texture decoders [3, 51, 62, 63, 83] significantly improves

the expressiveness of 3DMM. Some methods [13, 38] use linear 3DMM as a basis and learn animated displacement residual maps to compensate for shape or texture details. Nonlinear models are designed by jointly considering facial albedo, shape, normals, and displacement maps, as mentioned in [15, 39]. In addition, [70] relies on geometric consistency in multiple views and implicit neural rendering, but these are highly dependent on accurate camera pose.

Facial texture generation. Facial texture generation aims to create photo-realistic face images based on 3D faces. Initially, traditional rendering techniques project the vertex color of the mesh onto a 2D image plane to represent the texture. With the support of differentiable rendering techniques, several methods use self-supervised or weakly supervised learning to achieve the high-fidelity reconstruction of faces. For example, [9, 13] use mixed-level image information to achieve faithful face reconstruction and obtain more realistic facial textures, enabling the use of a large number of images in the wild. Some methods [53, 62, 63] use 2D UV texture representations, which help the neural networks obtain high-quality rendered images. Image translation-based methods [53] usually obtain coarse texture maps first and then map them to fine textures using the pix2pix method [65]. The texture decoder-based methods [4, 16, 36, 56, 62, 63] take advantage of StyleGAN2's [30] ability to generate high-resolution UV images. Then, the 3DMM matching algorithm is used to find the best latent code for reconstruction. Iterative strategies from coarse to fine used by both HRN [38] and NextFace [10] to get realistic results when reconstructing geometric details and textures. However, these iterative fitting methods are prone to overfitting on occlusions. FitMe [36] and Relightify [46] based on the face albedo dataset with separated lighting obtained from real world scans of AvatarMe++, respectively using the diffusion model [50] and GAN-tuning [49], realize extracting a face albedo map and lighting information from a single image. Furthermore, similar to our BFM-UV dataset production pipeline, FFHQ-UV [4] standardizes facial images and creates a high-resolution UV texture dataset using StyleFlow [1] and HIFI3D++ [6], reducing the production cost of the dataset while not retaining the identity of the original in-the-wild images.

Diffusion probabilistic models. The Diffusion Probabilistic Model (DM) is first proposed in [58] to generate images. It uses a markov diffusion chain to add gaussian noise to the input image iteratively and trains a denoising network to transform Gaussian noise to the desired images gradually. Recently, with the great success of DDPM [22] in generating high-quality images, DMs attract much attention in various generative fields, including image synthesis [23, 52] and audio synthesis [33, 44]. In particular, LDM [50] trains a denoising network based on attention mechanisms in the latent space to reduce computation costs. It demonstrates powerful multimodal controllable generation and has been widely applied in various areas, such as image editing [20, 26, 40, 79], 3D content generation [41, 48], brain signal visualization [59, 77], and video generation [2, 24, 78]. [72, 76] further demonstrate the high controllability of using images as prompts, inspiring our method to encode incomplete textures and feed them into the diffusion process by the cross-attention mechanism.

3. Methodology

The goal of our method is to obtain real face UV-texture from a single in-the-wild face image, which consists of two stages: creating a face UV-texture dataset (Sec. 3.1) and training a condition-guided latent diffusion model as the UV-texture generator (Sec. 3.2). During the inference process, we can provide incomplete facial UV texture to the UV-texture generator to obtain realistic and high-fidelity UV texture corresponding to the face images. These UV textures can be used in any BFM-based 3D face reconstruction method.

3.1. Dataset Creation

The dataset creation pipeline, shown in Fig. 2, consists of two steps: StyleGAN-based face image editing (Sec. 3.1.1) and UV-texture extraction (Sec. 3.1.2).

3.1.1 StyleGAN-Based Face Image Editing

In order to generate high-fidelity texture from an arbitrary face image, we initially produce a batch of three-view, unobstructed (free from occlusions such as glasses and hair) wild-face images. We leverage a pre-trained StyleGAN2 [30] on the FFHQ dataset [29] for random generation of these wild images and employ InterFaceGAN [55] for automated editing of image attributes within its $W+$ latent space [69]. We log the associated latent code within the $W+$ latent space for each face image to modify the semantic attributes correlated with the image. Given our objective to create textures that closely resemble the original image, we retain attributes related to lighting and expressions. The only attributes we want to remove are those related to hair, glasses, posture, etc. Considering the generation of texture details relies heavily on the foreground image, posture normalization becomes necessary. We detect the Euler angle with the pre-trained posture detection network and retain subject face images with pitch, yaw, and roll angles less than five degrees for further processing. This approach ensures the generation of forward-facing face images. As the generated images are synthesized and can be produced in bulk, we can thoroughly clean up facial images without worrying about dataset size limitations.

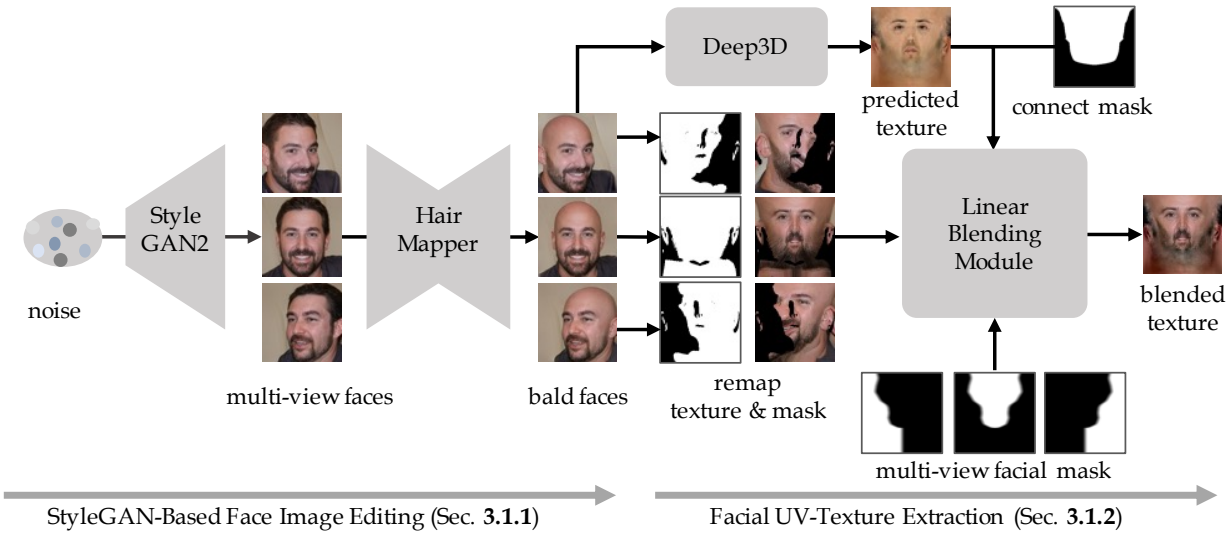

Figure 2. Our pipeline to generate a normalized texture UV-map from multi-view images. It consists of two parts: StyleGAN-based face image editing and face UV-texture extraction.

In alignment with InterFaceGAN, we employ SVM as a face attribute classifier to identify the associated attribute modification direction. This allows us to alter attributes along this direction and regulate the modification intensity via coefficients. By manipulating the attribute direction vectors related to the yaw angle, we create images of faces oriented toward the left and right. This process aids in accounting for texture augmentation in profile views of faces. Analogous to the process of posture normalization, we apply a glasses classifier to remove glasses. This classifier is trained to sift through face images produced by StyleGAN2, retaining only those images that do not feature glasses, thereby producing a pool of potential candidates. Regarding hair removal, we adopt the approach presented by HairMapper [68], which utilizes a fully connected network to discern the pathway for hair removal within the StyleGAN2's latent space. It then modifies the latent code to generate a new portrait seamlessly integrated with the original portrait through poisson blending. The latent code for the original image is acquired via e4e encoding [61]. Consequently, we can create bald portraits that accurately represent the images across three different views.

3.1.2 Face UV-Texture Extraction

In order to extract unwrapping UV texture from face images, we apply a classical rendering technique [73] to obtain authentic texture colors. It involves five major steps: (1) Use Microsoft's face reconstruction model Deep3D [9] trained on BFM [47] to extract accurate 3D face shape from a single image. (2) Calculate the normal vector's direction based on the face structure's vertex coordinates, which

Figure 3. Examples of the created BFM-UV dataset.

helps to identify the visible 3D vertex index. (3) Use the projection relationship between the 3D structure and the face image to identify the colors of all 3D vertices corresponding to the image pixels. (4) Unwrap the colors of the 3D vertices into a comprehensive UV map using predefined UV coordinates. (5) Derive the mask of the visible region of the UV map based on the visibility of the vertices and also extract reliable incomplete UV texture from the comprehensive UV map to ensure that the pixel colors correspond to each visible vertex.

We follow the above steps to extract incomplete textures from bald-face images in different views. This aims to compensate for the loss of facial texture information caused by a single perspective. Building on this, unlike FFHQ-UV [4], which utilizes an average texture, we extract a specific UV texture corresponding to the forward-facing image based on the linear texture basis of Deep3D as the template. By employing YUV spatial color matching [4] and pre-defined face visibility masks, we fuse the template and the incom-

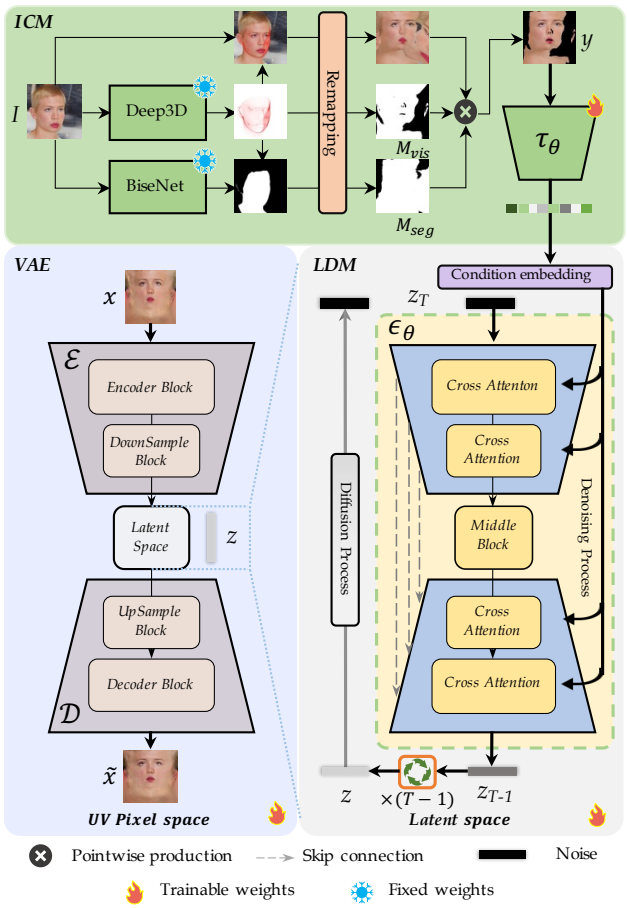

Figure 4. The details of UV-IDM. The upper part describes the process of obtaining identity conditions. The lower part is the UV texture generator based on the latent diffusion model. During training, the VAE is first trained separately, and then the ICM and LDM are jointly trained.

plete textures from the three angles using linear blending. This results in a complete and realistic UV-texture map, as shown in Fig. 3.

Overall, we first generate 280K images to ensure diversity in age and race. Then, based on the above steps, we also include 50K sets of normalized images provided by FFHQ-UV and ultimately generate nearly 80K high-quality BFM-based UV-texture maps with a resolution of 256×256. The collection of these data is called the BFM-UV dataset.

3.2. Identity-Conditioned Facial Texture Generator Based on a Latent Diffusion Model

In this section, we utilize the built BFM-UV dataset as the ground truth, with the aim of extracting a genuine face UV-texture map from an in-the-wild image. We introduce an end-to-end texture generator underpinned by an LDM.

We use LDM to learn the distribution $P(x)$ of UV texture data. Following [50], we train a variational autoencoder (VAE) [32] composed of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$ as a perceptual compression model, obtaining a low-dimensional latent space embedding $z = \mathcal{E}(x) \in \mathbb{R}^{32\times32\times4}$ that matches the UV texture data, by optimizing the loss term:

$$\mathcal{L}_{VAE} := \lambda_{kl}\mathcal{L}_{KL} + \lambda_{gan}\mathcal{L}_{GAN} + \lambda_{lpips}\mathcal{L}_{LPIPS}. \quad (1)$$

Here, $\mathcal{L}_{KL}$ denotes the KL divergence loss [32], $\mathcal{L}_{GAN}$ denotes the GAN adversarial loss [18], $\mathcal{L}_{LPIPS}$ denotes the LPIPS perceptual loss [80]. The loss weights λ_{kl}, λ_{gan} and λ_{lpips} control the relative contributions of the different loss terms during optimization. Compared with the high-dimensional UV pixel space, this latent space makes full use of the low-dimensional representation of the UV-texture space and is more suitable for training the likelihood-based generative model while effectively reducing memory consumption. Our texture VAE has the ability to capture the details of UV-texture maps, which also makes it possible to use it as a pre-train model in other 3D face reconstruction tasks. Subsequently, we train the diffusion model in this latent space. To overcome the challenge of preserving identity information during diffusion progress, we develop an Identity-Conditioned Module (ICM) that utilises information obtained from the origin images as conditions to guide the training of LDM, as shown in Fig. 4

Specifically, we consider the generation process a texture-completion task. Given a 2D facial image I as input, we obtain its incomplete texture by the ICM as identity guidance condition y (see Fig. 4). We initially employ Deep3D to reconstruct the aligned facial image and acquire the 3D vertices. Then, we separately remap the facial mask M_{seg} and the vertex mask M_{vis} in the UV space. This is accomplished based on the facial regions partitioned by the facial segmentation network BiSeNet [74] and the vertex visibility determined by each vertice's normal vector. The former isolates facial regions from hair and glasses, thereby mitigating the impact of occlusions. Coupled with the composite texture mentioned in step 5 in Sec. 3.1.2, we ultimately obtain the incomplete texture as the condition y.

We encode the incomplete texture y through an embedding network τ_θ to obtain condition embeddings and then inject them into different levels of the noise prediction network ϵ_θ within the LDM through the cross-attention mechanism, similar to handling image prompts [72, 76]. Finally, we train our LDM based on the incomplete UV condition y and the complete UV x via the following loss:

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, \tau_\theta(y)\right)\|_2^2\right], \quad (2)$$

where t is the timestep uniformly sampled from $\{1, ..., T\}$; during training, the denoised variant z_t is obtained based on t and z; and the final sampled latent can be decoded into the UV-texture space through $\mathcal{D}$.

Figure 5. Comparison of rendered images using different methods.

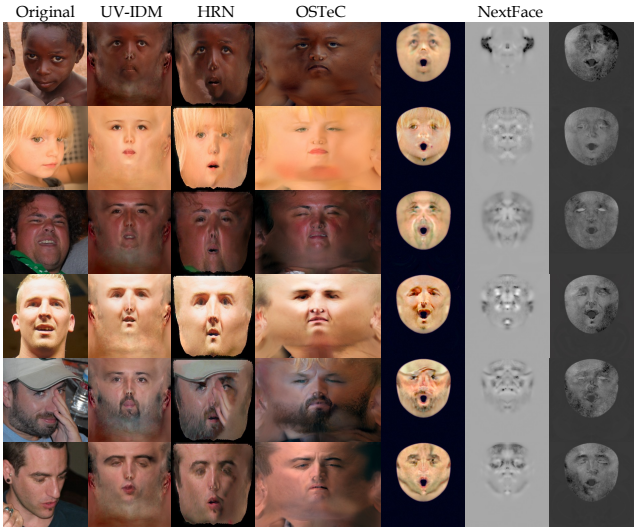

Figure 6. Examples of texture generation from in-the-wild images using OSTeC, HRN, NextFace and our method.

By extracting incomplete textures in the identity condition module, we can obtain as much information as possible from in-the-wild images, such as side faces with large poses, cheeks occluded by the hair, etc. During inference, our identity condition module can obtain facial visibility masks M_{seg} through BiseNet to ensure that hair and glasses areas are excluded from the incomplete textures. Our UV-IDM has sufficient capability to restore accurate facial details, ensuring higher accuracy and personalization of the generated UV texture maps.

4. Experiments

Implementation details. Our training on texture generation consists of two parts: training the VAE and training the LDM on our BFM-UV dataset. The training of the VAE is performed by selecting the UV texture maps of the faces as self-supervised learning. We use Adam optimizer [31] with $\beta_1 = 0.5$ and $\beta_2 = 0.9$ to optimize our model with a learning rate of $\eta = 5.76 \times 10^{-4}$ and the loss weights $\lambda_{KL} = 10^{-6}$, $\lambda_{GAN} = 0.5$ and $\lambda_{LPIPS} = 1$. To train the LDM, we use our BFM-UV texture maps as the ground truth images and employ the identity condition module to extract incomplete textures from the original multi-view facial images as conditions. To further mitigate the impact of hair on facial texture generation, our approach uses paired training data consisting of the original images with hair and the corresponding textures after hair removal as inputs for network training. We use AdamW optimizer [45] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay $\mu = 0.01$ and learning rate $\eta = 5.76 \times 10^{-4}$ to optimize the LDM. Training VAE takes three days and LDM five days on 8 40 GB NVIDIA A100-SXM4 GPUs.

Benchmark datasets. To evaluate our method and ensure a balanced distribution of attributes such as age and gender in the validation set, we randomly sample 1195 and 2250 face images of different identities from the CelebAMask-HQ [37] and FFHQ [29] datasets, respectively. These images are never seen during our training process, but the face poses in these datasets are relatively standard with rare occlusions. Furthermore, to assess the robustness of our method to occlusions and pose variations, we also randomly select 496 images from the widely used challenging face dataset AFLW2000 [82] as test data.

Evaluation metrics. We adopt multiple metrics to evaluate the quality of UV-texture maps by rendering them back to the aligned facial images used as ground truth images. We use Learning Perception Image Patch Similarity (LPIPS) [80] to measure reconstruction accuracy and use Fréchet Inception Distance (FID) [21] to measure the visual quality. We use cosine similarity (CSIM) [75, 78] of face identity embeddings to measure the ability of identity retention.

4.1. Results

Qualitative analysis. We show the qualitative comparison with state-of-the-art (SOTA) methods in Fig. 5, where the examples are randomly selected from the FFHQ, CelebAMask-HQ, and AFLW2000 datasets. We use UV-IDM to generate textures from wild images and replace the original textures of Deep3D [9] and HRN [38], denoted as D-UV-IDM and H-UV-IDM respectively. It can be observed that, despite the careful optimization of Deep3D, the final texture quality remains limited by the linear base representation ability, causing a blurry texture and neglecting

Figure 7. Texture generation methods based on iterative optimization may experience overfitting, such as FFHQ-UV, OSTeC, NextFace, and HRN, which may result in occlusion (hat) merging into the texture to some extent, while D-UV-IDM exhibits robustness to occlusion.

Table 1. Quantitative comparison of rendered face quality.

Method	FFHQ			CelebAMask-HQ			AFLW2000		
	LPIPS↓	FID↓	CSIM↑	LPIPS↓	FID↓	CSIM↑	LPIPS↓	FID↓	CSIM↑
HRN	**0.1484**	27.15	**0.9518**	**0.1433**	31.43	**0.9573**	**0.1536**	67.79	**0.9361**
H-UV-IDM	0.1527	**23.91**	0.9457	0.1474	**26.90**	0.9502	0.1635	**57.36**	0.9358
Deep3D	0.1638	25.63	0.9351	0.1578	28.71	0.9424	**0.1615**	62.14	0.9226
D-UV-IDM	**0.1575**	**22.65**	**0.9428**	**0.1546**	**24.92**	**0.9501**	0.1651	**57.15**	**0.9346**

the face's high-frequency features. The GAN-based method FFHQ-UV [4] tends to produce smooth results, unable to capture intricate facial details and lacking realism, such as the cheekbones and eyebrows of the female face in the first row. HRN is a multi-stage method that jointly infers and optimizes geometry and texture from coarse to fine. To minimize the impact of geometric optimization on texture rendering, we set the number of optimization iterations of HRN to 0. Even so, the inference stage of HRN still incorporates occlusions such as hair into the geometry and texture (e.g., the hair in the first, second, and fourth rows, and the microphone in the third row). However, compared to HRN, H-UV-IDM can better capture high-frequency details, such as the highlights on the female face in the fourth row without the need for iteration.

In Fig. 6 and Fig. 7, we qualitatively compare our method against other texture generation methods [10, 16, 38] based on iterative optimization under challenging scenarios including pose, skin tone, lighting and occlusion (the number of optimization iterations for HRN is set to 50). Our method achieves high-quality results in these scenes. For occlusions such as hair, hats and glasses, our method extracts conditional inputs from the incomplete texture to the texture generator, effectively reducing texture redundancy caused by face occlusion. Even for gesture occlusions unseen in the training dataset, our method can still avoid recovering them into the texture. HRN and NextFace [10] fit the hand into the texture, shown in the fifth row of Fig. 6. Additionally, due to the adversarial behaviour of the joint optimization process between geometry and texture, HRN fits the background into the texture, as depicted by the presence of the grey region and black artifact in the first and sixth rows of Fig. 6. This shows UV-IDM's robustness to occlusion and its ability to generate results that better match digital assets.

Quantitative analysis. We compare quantitatively with SOTA methods in Tab. 1. Since FFHQ-UV employs a distinct 3DMM and alignment approach, and iterative fitting-based methods such as OSTeC [16] and NextFace are prone to overfitting occlusions, we exclude these methods from our quantitative experiments. Because of the different ways of cropping and reconstructing 3D facial shapes, we split D-UV-IDM and H-UV-IDM into two groups so that we can compare them with Deep3D and HRN. The D-UV-IDM surpasses Deep3D in terms of LPIPS, FID, and CSIM on the FFHQ and CelebAMask-HQ datasets, showcasing superior reconstruction, generation, and identity preservation abilities. Likewise, we set the number of optimization iterations for HRN to 0. The AFLW dataset comprises numerous facial images in complex scenarios, necessitating greater capabilities from texture generators. In this dataset, the LPIPS of D-UV-IDM is slightly lower than Deep3D, but it has advantages in rendering quality and identity consistency. The LPIPS and CSIM scores of HRN on the three datasets surpass those of H-UV-IDM. Still, due to the inherent overfitting issue of HRN, this result is reasonable. Nonetheless, when comparing the FID scores, H-UV-IDM performs better in terms of rendering quality.

Tab. 3 demonstrates the superiority of our method in terms of time complexity. The inference time of D-UV-IDM or H-UV-IDM using a P40 GPU is merely 6 seconds, much faster than iterative GAN-based methods (e.g., FFHQ-UV, OSTeC, NextFace, and HRN with 50 steps of optimization). This is owing to the powerful generalization ability of the latent diffusion model, with which our method no longer requires hundreds of iterative optimizations, but only a single inference process to achieve similar generative results. Additionally, we employ the DDIM acceleration technique to reduce the denoising steps of the diffusion model to 50.

Table 2. Ablation study table on D-UV-IDM with different conditions.

Condition	FFHQ			CelebAMask-HQ			AFLW2000		
	LPIPS↓	FID↓	CSIM↑	LPIPS↓	FID↓	CSIM↑	LPIPS↓	FID↓	CSIM↑
Origin Image	0.1625	23.39	0.9417	0.1581	**23.74**	**0.9512**	0.1703	**56.80**	0.9233
Incomplete UV	**0.1575**	**22.65**	**0.9428**	**0.1546**	24.92	0.9501	**0.1651**	57.15	**0.9346**

Table 3. Comparison of inference times among different methods. (a) D-UV-IDM. (b) H-UV-IDM. (c) FFHQ-UV. (d) OSTeC. (e) NextFace. (f) HRN.

Methods	(a)	(b)	(c)	(d)	(e)	(f)
time (s)	6	6	150	800	160	18

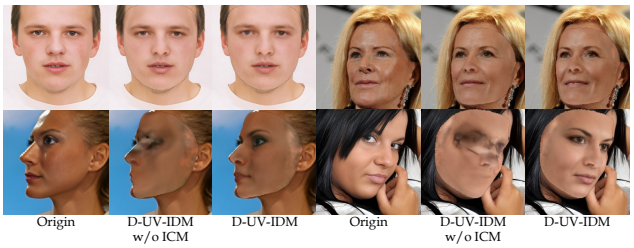

Figure 8. The visual examples of our ablation analysis.

4.2. Ablation Study

We emphasize the significance of our identity condition module by comparing the results of inputting the original image as a condition. Specifically, we use the same embedding network to encode the original image into condition embeddings to guide the diffusion process through the cross-attention mechanism and training for the same duration. The visualization indicates that, without exaggerated posture and obvious occlusion, inputting the original image as a condition can also generate realistic and detailed texture. As shown in Fig. 8, the details of women in the upper-right face and the skin colors of the men are similarly good, achieved by both methods. Nevertheless, when there is an exaggerated pose and occlusion (such as hair), as shown in the bottom row, UV-IDM effectively restores the area occluded by the hair and shows stronger robustness to pose variation. From the results on different datasets in Tab. 2, it can be seen that in the CelebMask-HQ dataset, the rendered image generated by feeding the original image as the condition even has better FID and CSIM scores than using the incomplete UV texture as the condition. However, based on the results of the AFLW dataset, incomplete UV textures as the conditions achieve comparable quality and better identity consistency, demonstrating excellent performance on faces taken in complex environments.

5. Discussion and Conclusion

The use of synthetic data to overcome the obstacle of acquiring insufficient real-world data is a widely adopted approach in the academic community [64, 67]. However, this approach imposes greater demands on the accuracy, diversity, generalization, and differentiation from real data to synthetic data. This work, similar to FFHQ-UV, seeks to address the difficulty of obtaining texture data in the real world and minimize data acquisition costs by leveraging the SOTA face image generation model StyleGAN2 [30]. Nonetheless, it is important to note that the generated data may inherit the inductive bias of the FFHQ dataset [29]. Moreover, edits based on StyleGAN2 may result in the loss of identity information. although our primary focus is on frontal images within the multi-view face images, side images are utilized to add profile details. In the future, we will consider using 3D-aware GAN [7, 8, 27] to generate texture data demonstrating better face identity performance and structural consistency. We will also explore the integration of geometric generation and light separation to produce usable digital assets directly. Incorporating existing visual language models (VLMs) [28, 42, 43, 66, 71] to generate customized digital assets is also part of our consideration.

Conclusion. In this paper, we develop an identity-conditioned, LDM-based end-to-end high-quality texture generator for BFM that can extract valuable information from in-the-wild images and generate high-fidelity, identity-consistent texture maps while maintaining robustness to complex factors such as large poses and hair occlusions. Additionally, by leveraging the editability of StyleGAN2, we build a dataset with (in-the-wild image, hair removal image, UV texture) triples, generating over 80K high-fidelity UV texture maps, allowing more extensive research in this area.

Acknowledgements. The work was supported by the National Key Research and Development Program of China (2023YFC3300029), Zhejiang Provincial Natural Science Foundation of China(LD24F020007), Beijing Natural Science Foundation (L223024), National Natural Science Foundation of China (62076016), “One Thousand Plan” projects in Jiangxi Province (Jxsg2023102268), Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Z231100005923035).

References

[1] Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM TOG*, 2021. 2, 3

[2] Anonymous. Tokenflow: Consistent diffusion features for consistent video editing. In *ICLR*, 2023. 3

[3] Timur Bagautdinov, Chenglei Wu, Jason Saragih, Pascal Fua, and Yaser Sheikh. Modeling facial geometry using compositional vaes. In *CVPR*, 2018. 2

[4] Haoran Bai, Di Kang, Haoxian Zhang, Jinshan Pan, and Linchao Bao. Ffhq-uv: Normalized facial uv-texture dataset for 3d face reconstruction. *arXiv:2211.13874*, 2022. 1, 2, 3, 4, 7

[5] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999. 1, 2

[6] Zenghao Chai, Haoxian Zhang, Jing Ren, Di Kang, Zhengzhuo Xu, Xuefei Zhe, Chun Yuan, and Linchao Bao. Realy: Rethinking the evaluation of 3d face reconstruction. In *ECCV*, 2022. 2, 3

[7] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini de Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, 2022. 8

[8] Boyang Deng, Yifan Wang, and Gordon Wetzstein. Lumigan: Unconditional generation of relightable 3d human faces. *arXiv:2304.13153*, 2023. 8

[9] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPR Workshops*, 2020. 1, 2, 3, 4, 6

[10] A. Dib, G. Bharaj, J. Ahn, C. Thébault, P. Gosselin, M. Romeo, and L. Chevallier. Practical face reconstruction via differentiable ray tracing. *CGF*, 2021. 1, 3, 7

[11] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3d morphable face models—past, present, and future. *ACM TOG*, 2020. 1, 2

[12] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. 2

[13] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM TOG*, 2021. 1, 2, 3

[14] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *CVPR*, 2019. 1

[15] Baris Gecer, Alexandros Lattas, Stylianos Ploumpis, Jiankang Deng, Athanasios Papaioannou, Stylianos Moschoglou, and Stefanos Zafeiriou. Synthesizing coupled 3d face modalities by trunk-branch generative adversarial networks. In *ECCV*, 2020. 3

[16] Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. Ostec: One-shot texture completion. In *CVPR*, 2021. 1, 3, 7

[17] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview face capture using polarized spherical gradient illumination. *ACM TOG*, 2011. 2

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 5

[19] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *ECCV*, 2020. 2

[20] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023. 3

[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 *NeurIPS*, 2017. 6

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[23] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. In *JMLR*, 2022. 3

[24] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 3

[25] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. *ACM TOG*, 2017. 2

[26] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Shifeng Chen, and Liangliang Cao. Diffusion model-based image editing: A survey. *arXiv:2402.17525*, 2024. 3

[27] Kaiwen Jiang, Shu-Yu Chen, Hongbo Fu, and Lin Gao. Nerffacelighting: Implicit and disentangled face lighting representation leveraging generative prior in neural radiance fields. *ACM TOG*, 2023. 8

[28] Yiqiao Jin, Minje Choi, Gaurav Verma, Jindong Wang, and Srijan Kumar. Mm-soc: Benchmarking multimodal large language models in social media platforms. *arXiv:2402.14154*, 2024. 8

[29] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 3, 6, 8

[30] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2, 3, 8

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[32] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, 2013. 5

[33] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *ICLR*, 2020. 3

[34] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet

Ghosh, and Stefanos Zafeiriou. Avatarme: Realistically renderable 3d facial reconstruction" in-the-wild". In *CVPR*, 2020. 1, 2

[35] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos Zafeiriou. Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans. *IEEE TPAMI*, 2022.

[36] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. Fitme: Deep photorealistic 3d morphable model avatars. In *CVPR*, 2023. 1, 2, 3

[37] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 6

[38] Biwen Lei, Jianqiang Ren, Mengyang Feng, Miaomiao Cui, and Xuansong Xie. A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images. *arXiv:2302.14434*, 2023. 2, 3, 6, 7

[39] Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Prasad, Bipin Kishore, Jun Xing, and Hao Li. Learning formation of physically-based face attributes. In *CVPR*, 2020. 3

[40] Shanglin Li, Bohan Zeng, Yutang Feng, Sicheng Gao, Xuhui Liu, Jiaming Liu, Li Lin, Xu Tang, Yao Hu, Jianzhuang Liu, et al. Zone: Zero-shot instruction-guided local editing. *arXiv:2312.16794*, 2023. 3

[41] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, 2023. 3

[42] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv:2310.03744*, 2023. 8

[43] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 8

[44] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, Peng Liu, and Zhou Zhao. Diffsinger: Diffusion acoustic model for singing voice synthesis. *arXiv:2105.02446*, 2021. 3

[45] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv:1711.05101*, 2017. 6

[46] FoivosParaperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, and Stefanos Zafeiriou. Relightify: Relightable 3d faces from a single image via diffusion models. In *CVPR*, 2023. 1, 2, 3

[47] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *AVSS*, 2009. 2, 4

[48] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 3

[49] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM TOG*, 2021. 3

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 5

[51] Zeyu Ruan, Changqing Zou, Longhai Wu, Gangshan Wu, and Limin Wang. Sadrnet: Self-aligned dual face regression networks for robust 3d dense face alignment and reconstruction. *IEEE TIP*, 2021. 2

[52] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *SIGGRAPH*, 2022. 3

[53] Shunsuke Saito, Lingyu Wei, Liwen Hu, Koki Nagano, and Hao Li. Photorealistic facial texture inference using deep neural networks. In *CVPR*, 2017. 3

[54] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *CVPR*, 2017. 1

[55] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020. 2, 3

[56] Ron Slossberg, Ibrahim Jubran, and Ron Kimmel. Unsupervised high-fidelity facial texture generation and reconstruction. In *ECCV*, 2022. 3

[57] William A. P. Smith, Alassane Seck, Hannah Dee, Bernard Tiddeman, Joshua B. Tenenbaum, and Bernhard Egger. A morphable face albedo model. In *CVPR*, 2020. 1

[58] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 3

[59] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023. 3

[60] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCV*, 2017. 2

[61] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM TOG*, 2021. 4

[62] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *CVPR*, 2018. 1, 2, 3

[63] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *CVPR*, 2019. 1, 2, 3

[64] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, Baining Guo, and Microsoft Research. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *CVPR*, 2023. 8

[65] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 3

[66] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming

Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. *arXiv:2311.03079*, 2023. 8

[67] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J. Cashman, and Jamie Shotton. Fake it till you make it: Face analysis in the wild using synthetic data alone. In *ICCV*, 2021. 8

[68] Yiqian Wu, Yong-Liang Yang, and Xiaogang Jin. Hairmapper: Removing hair from portraits using gans. In *CVPR*, 2022. 2, 4

[69] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *CVPR*, 2021. 3

[70] Yunze Xiao, Hao Zhu, Haotian Yang, Zhengyu Diao, Xiangju Lu, and Xun Cao. Detailed facial geometry recovery from multi-view images by learning an implicit function. In *AAAI*, 2022. 3

[71] Yijia Xiao, Yiqiao Jin, Yushi Bai, Yue Wu, Xianjun Yang, Xiao Luo, Wenchao Yu, Xujiang Zhao, Yanchi Liu, Haifeng Chen, et al. Large language models can be good privacy protection learners. *arXiv:2310.02469*, 2023. 8

[72] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arxiv:2308.06721*, 2023. 3, 5

[73] Xiangnan Yin, Di Huang, Zehua Fu, Yunhong Wang, and Liming Chen. Weakly-supervised photo-realistic texture generation for 3d face reconstruction. In *FG*, 2023. 1, 4

[74] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, 2018. 5

[75] Bohan Zeng, Boyu Liu, Hong Li, Xuhui Liu, Jianzhuang Liu, Dapeng Chen, Wei Peng, and Baochang Zhang. Fnevr: Neural volume rendering for face animation. In *NeurIPS*, 2022. 2, 6

[76] Bohan Zeng, Shanglin Li, Yutang Feng, Hong Li, Sicheng Gao, Jiaming Liu, Huaxia Li, Xu Tang, Jianzhuang Liu, and Baochang Zhang. Ipdreamer: Appearance-controllable 3d object generation with image prompts. *arXiv:2310.05375*, 2023. 3, 5

[77] Bohan Zeng, Shanglin Li, Xuhui Liu, Sicheng Gao, Xiaolong Jiang, Xu Tang, Yao Hu, Jianzhuang Liu, and Baochang Zhang. Controllable mind visual diffusion model. *arXiv:2305.10135*, 2023. 3

[78] Bohan Zeng, Xuhui Liu, Sicheng Gao, Boyu Liu, Hong Li, Jianzhuang Liu, and Baochang Zhang. Face animation with an attribute-guided diffusion model. In *CVPRW*, 2023. 3, 6

[79] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 3

[80] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5, 6

[81] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *CVPR*, 2015. 2

[82] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *CVPR*, 2016. 6

[83] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z. Li. Face alignment in full pose range: A 3d total solution. *IEEE TPAMI*, 2017. 2

[84] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. *CGF*, 2018. 2