Physically Based Facial Texture Generation In The Wild

Chi Wang*, Junming Huang*, Rong Zhang*, Qi Wang, Haotian Yang, Pengfei Wan, Haibin Huang, Chongyang Ma, and Weiwei Xu†, *Member, IEEE*

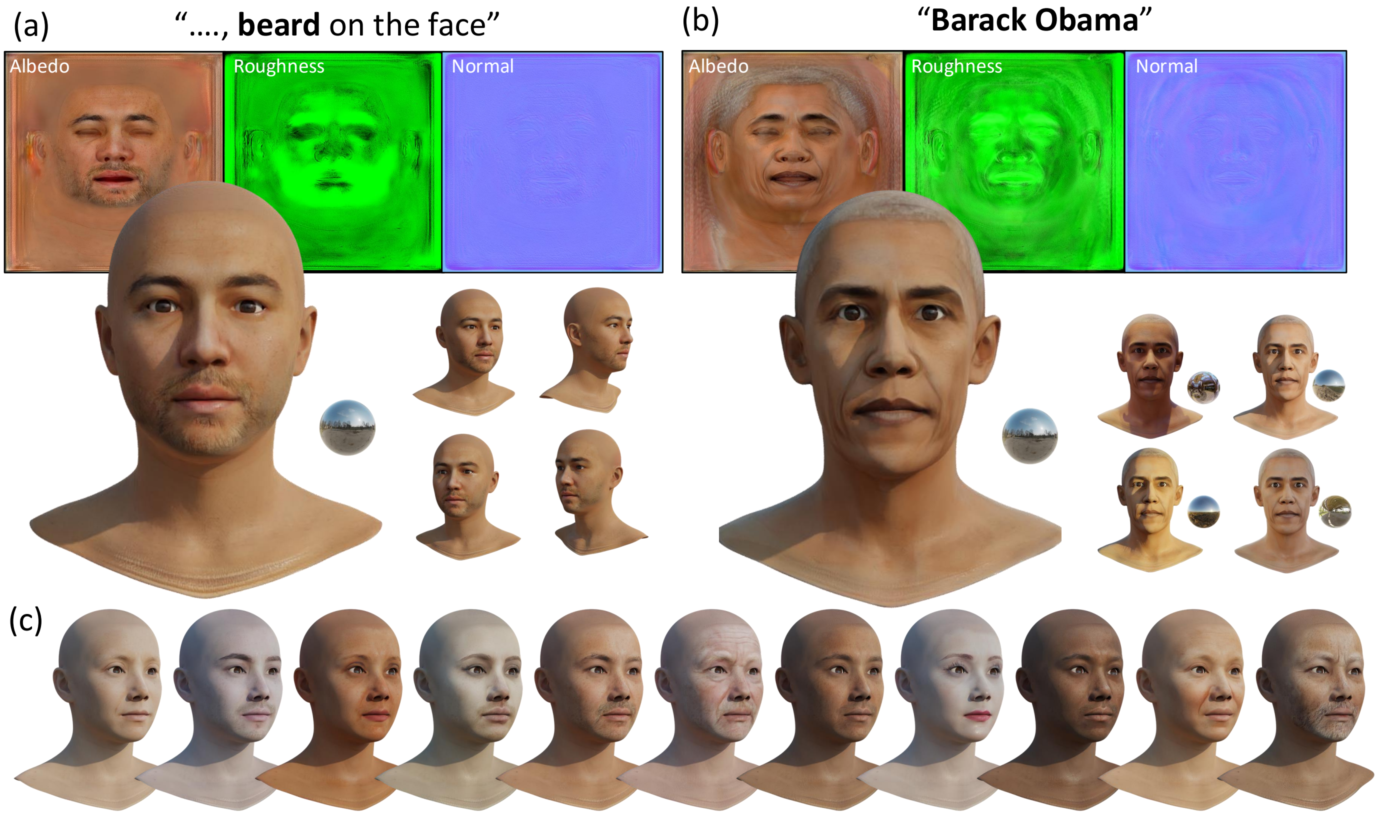


Fig. 1: Our method can faithfully generate a variety of PBR facial textures from text prompts for photo-realistic rendering. From left to right: (a) multi-view rendering results using our generated PBR texture from a prompt "beard". (b) relighting results using our generated PBR texture of "Barack Obama". (c) unconditional generation results.

***Abstract*—Automatic 3D facial texture generation has gained significant interest recently. However, existing approaches may lack compatibility with the widely used physically based rendering (PBR) pipeline or rely on 3D data captured by sophisticated systems such as Light Stage. In this paper, we propose a multi-stage framework to achieve text-driven physically based facial texture generation in the wild, which eliminates the reliance on expensive, controlled capture environments. It is based on FFHQ-UV to pave the way between the normalized UV texture space and facial images captured in unconstrained real-world settings and remove the influence of the background or hair in natural images on PBR texture generation. Specifically, we first integrate differentiable rendering techniques and carefully crafted texture disentanglement regularization to train a generative adversarial network for efficient PBR texture sampling. Then, the latent space of the network is aligned with the text embedding space for flexible text-guided generation. Besides, we design an edge-aware Score Distillation Sampling (EASDS) loss and introduce an EASDS-based PBR texture boosting scheme to achieve more diverse generation and efficient SDS optimization. Experiments demonstrate that our method outperforms existing PBR texture generation methods.**

***Index Terms*—Multimodal synthesis; Color, shading, shadowing, and texture.**

C. Wang, J Huang, Q Wang, W Xu are with Zhejiang University, Hangzhou, Zhejiang, China.
R. Zhang is with Zhejiang Key Lab of Big Data and Future E-Commerce Technology, Zhejiang Gongshang University, Hangzhou, Zhejiang, China.
H Huang, C Ma are with Kwai Inc, San Francisco, CA, US.
H Yang, P Wan are with Kuaishou Technology, Beijing, China.
* Equal contribution.
† Corresponding author: Weiwei Xu.
Manuscript received October 17, 2024; revised April 22, 2025; accepted June 1, 2025.

I. Introduction

RECENTLY, 3D facial texture generation has gained widespread attention for its potential to boost applications in augmented reality (AR), virtual reality (VR), gaming, and films. The traditional pipeline of creating textures typically

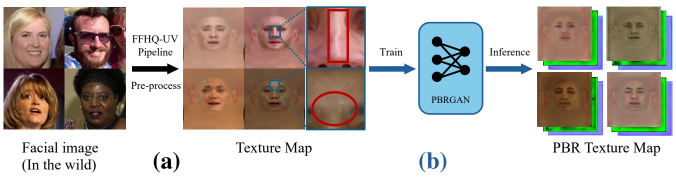


Fig. 2: The illustration of the data preparation (a) and the training pipeline (b) of PBRGAN. The entangled specular highlights of FFHQ-UV texture are labeled with blue boxes and enlarged for details. The red box highlights the region.

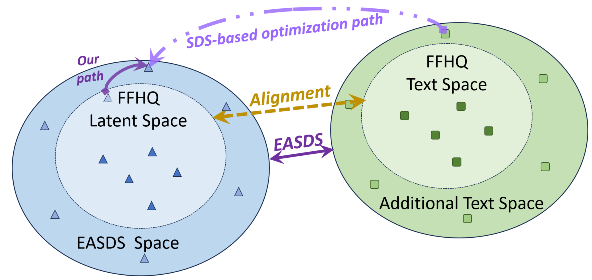


Fig. 3: The key stages of our framework. The latent space and text space are aligned for text-guided generation. For uncommon prompts outside the FFHQ text space, PBRGAN can provide better initialization for SDS optimization, shortening the SDS-based optimization path.

involves UV mapping, material production, or texture synthesis. With the advancement of large-scale pre-trained models, such as Stable Diffusion [1] and the Contrastive Language-Image Pre-Training (CLIP) model, textures can be generated through a more flexible and efficient pathway, *i.e.* , text-driven texture generation. Pioneers have made effective attempts to obtain facial textures under the guidance of text prompts [2], [3]. However, these methods encounter challenges in generating disentangled facial textures for physically based rendering (PBR), which can greatly broaden the application of generative models. Typical PBR facial texture consists of albedo, roughness, and normal maps, as illustrated in Fig. 1. They usually do not have hairs to facilitate subsequent hairstyle editing applications.

With the help of the differentiable rendering technique, recent researches are capable of generating PBR textures using CLIP loss [4], [5] or Score Distillation Sampling (SDS) [6], [7]. Nevertheless, they require extra efforts to optimize for the textures given different 3D objects or text prompts. To address this issue, DreamFace [8] proposes a diffusion-based generative model in both the image and latent space, yielding high-quality albedo texture maps. This method is supervised with explicit PBR data captured by a light stage [9]. However, light stages suffer from drawbacks such as high equipment costs, high technical complexity, time-consuming, complex post-processing, *etc*. It is still challenging to generate facial textures using images in the wild.

In this paper, we investigate how to generate PBR facial textures in the wild, *i.e.* using facial images captured in unconstrained real-world settings. It is based on a recent contribution FFHQ-UV [10], which efficiently converts natural facial images into the normalized UV texture space needed for PBR facial textures. This conversion diminishes the effect of non-facial areas, such as the background and hair, on the generation of PBR facial texture. Nevertheless, because FFHQ-UV neither produces normal and roughness maps nor separates facial UV texture from illumination information as shown in Fig. 2(a), these entangled textures are unsuitable for direct use in PBR and may result in visible artifacts, for instance, light spots (*Texture Map* in Fig. 2).

Our approach, named PBRGAN, aims to design a generative adversarial network (GAN) to produce PBR facial textures, aided by a training dataset created using FFHQ-UV textures. Specifically, we synthesize a multi-view, non-hair, and non-background facial image training dataset by directly querying the color of each vertex or pixel with UV coordinates from the FFHQ-UV facial textures during rendering. Consequently, utilizing a differentiable rendering module, the GAN can be trained to align the facial images rendered with the generated PBR textures to the synthesized dataset. This process does not require supervision from the collected PBR textures but instead relies solely on wild facial images (Fig. 2). We refrain from directly using diffusion models in this study owing to their slow training speed. Additionally, we demonstrate that initializing the PBR textures with GAN can accelerate optimization with Score Distillation Sampling (SDS) loss.

PBRGAN comprises three stages to achieve text-driven, diverse facial PBR texture generation. First, we construct a GAN model that integrates differentiable rendering [11] and a self-supervised training scheme to disentangle the FFHQ-UV textures from wild facial images into PBR textures. Next, the model is enhanced with cross-modal generation capability via CLIP-based text embedding and latent space alignment, which enables PBRGAN to generate reasonable results under the guidance of text prompts. Furthermore, we propose an edge-aware SDS (EASDS) loss with explicit facial feature-line prior and an optional EASDS-based boosting stage to distill relevant knowledge of the pre-trained diffusion model for more diverse generation results. We provide better initial textures for SDS by leveraging the PBRGAN generated textures. As illustrated in Fig. 3, the EASDS effectively shortens the SDS optimization path.

In summary, our contributions are as follows:

- We propose PBRGAN that can bootstrap from FFHQ-UV textures for facial PBR texture generation in the wild. This model is trained with carefully designed training strategy and regularization terms without requiring supervision from explicitly disentangled PBR data.
- We present a three-stage framework to generate diverse PBR textures under the guidance of text prompts. In addition, we design a novel ControlNet-based edge-aware SDS loss to preserve the key facial structures.
- Our framework can generate high-quality facial PBR texture maps with the highest $TOPIQ_{face}$ among state-of-the-art methods and our user study score. Given a specific text prompt, the response time of our method is lower than state-of-the-art approaches based on diffusion models.

II. Related Works

Generative models, such as GANs and diffusion models, have been developed rapidly in recent years [1], [12]–[20]. In this paper, we review researches on 3D texture generation mostly related to our work.

a) 3D-aware image synthesis aims to map random vectors to 3D representations for multi-view consistent image synthesis. Visual object network [21] proposes a fully differentiable 3D-aware generative model for image and shape synthesis with a disentangled 3D representation. Liao et al. [22] consider the generative process into a 3D content creation stage and a 2D rendering stage, yielding a 3D controllable image synthesis model. However, 3D data collections are indispensable for these methods.

Follow-up works introduce neural representations and leverage 2D data to address the dataset limitation [23]–[27]. Pi-GAN [28] and ShadeGAN [29] represent scenes as view-consistent neural radiance fields and map 3D coordinates to pixel values as a 3D prior. EG3D [30] and Next3D [31] adopt feature triplanes for more efficient 3D representation. Yuan et al. [32]leverage the inherent properties of EG3D's latent space to design a discriminator and a background depth regularization, enabling 3D GAN inversion through geometry and occlusion-aware encoding. G-NeRF [33] enhances geometry priors by employing a geometry-guided multi-view synthesis approach followed by depth-aware training. GSGAN [34] utilizes a hierarchical multi-scale Gaussian representation to regularize the position and scale of generated Gaussians. Another series of approaches is dedicated to improving the controllability of 3D-aware generative models [35]–[38]. CGOF++ [39] presents a NeRF-based framework for controllable 3D face synthesis by integrating explicit 3D priors from 3D Morphable Models (3DMM). SwitchLight [40] introduces a physics-guided architecture with a pre-training framework for human portrait relighting by enhancing the physical reflectance model and adopting self-supervised learning. Nonetheless, these methods only generate 2D images and are difficult to use in the classic 3D content creation pipeline.

b) Text-to-3D generation aims to directly generate 3D content from textual descriptions, which further improve the flexibility compared to Image-to-3D methods [41]–[48]. Recently, with the advent of large-scale cross-modal pre-training models, text-guided 3D generative models [2], [3], [41], [49], [50] have become a frontier research topic. Many researchers utilize the CLIP model [51] to ensure the generated 3D contents match the text prompts. Text2Mesh [52] optimizes the color attributes and geometric displacements of individual vertices simultaneously. CLIP-NeRF [53] presents a NeRF-based framework to provide users with flexible control over 3D content using either a text prompt or an exemplar image. However, these methods may fail to generate high-quality results.

Recently, the powerful generative capabilities of diffusion models have been integrated into text-to-3D generation [54]–[58]. DreamFusion [6] adopts pre-trained text-to-image diffusion models as priors and generates realistic 3D models via SDS optimization. ProlificDreamer [59] further extends SDS to Variational Score Distillation, which optimizes the distribution of 3D scenes. HeadSculpt [60] employs the gradients combination of Stable Diffusion [1] and InstructPix2Pix [61] to craft 3D head avatars in a coarse-to-fine manner. These 3D generation techniques perform joint optimization of geometry and texture, considering them as a unified entity. Meanwhile, some approaches decouple geometry from texture, allowing independent refinement of both components, which can enhance generalization and adaptability. FaceG2E [62] leverages pre-trained image-edit diffusion models to construct a geometry-texture decoupled pipeline for text-guided 3D face generation and editing. TEXTure [63] iteratively renders the object from different viewpoints and applies a depth-to-image 2D diffusion model to generate a complete unwrapped UV texture map. However, the above methods generate textures in entangled representations, which are incompatible with the PBR pipeline.

c) PBR texture generation plays an essential role in achieving photorealistic results within existing 3D content production pipelines. Some methods [64]–[67] employ 3DMM [41], [68]–[70] to generate physical-based texture maps with the guidance of reference images or random noises. Relightify [71] utilizes a high-quality UV dataset of facial reflectance components to train an unconditional diffusion model for relightable 3D face reconstruction from images. MoSAR [72] employs a semi-supervised training scheme that leverages both light stage and in-the-wild datasets to generate 3D avatars from monocular images. In recent contributions, CLIP loss and SDS loss have been applied to text-driven PBR texture generation, such as CLIP-mesh [4], TANGO [5], and Fantasia3D [7]. However, these models must be retrained for each 3D object or each text prompt, which is time-consuming.

DreamFace [8] presents a diffusion model-based scheme to generate personalized 3D faces with text guidance. It can achieve physically-based rendering by decomposing compact latent space into diffuse albedo, specular intensity, and normal maps. The decomposing model relies on a captured high-quality PBR texture dataset, which is usually collected for commercial purposes and has not been released to the public. Besides, the inference efficiency of DreamFace is limited by the diffusion process. ID2Reflectance [73] proposes a framework for reconstructing high-quality facial reflectance maps from a single image using limited reflectance data for training. It leverages pretraining multi-domain facial feature codebooks and identity-conditioned swapping module to directly estimate the reflectance maps from a given image. However, ID2Reflectance still requires ground-truth PBR texture data for training and does not support text-driven PBR texture generation. In this paper, we expand the $\mathcal{W}_+$ feature space of StyleGAN to support text-to-3D PBR texture generation with no captured datasets. It will benefit from both the efficiency of GANs and the controllability of text-based generation.

III. Our Method

In this section, we present PBRGAN, a framework for text-driven physically based facial texture generation in the wild. Given the geometry of a 3D face, PBRGAN aims to generate the face texture maps under the guidance of text prompts. The textures are decomposed into three components, *i.e.* , albedo, normal, and roughness maps for physically based rendering without the supervision of explicitly disentangled data. The whole pipeline comprises three stages: (1) Disentanglement (Sec. III-A). We train a PBR StyleGAN to bootstrap

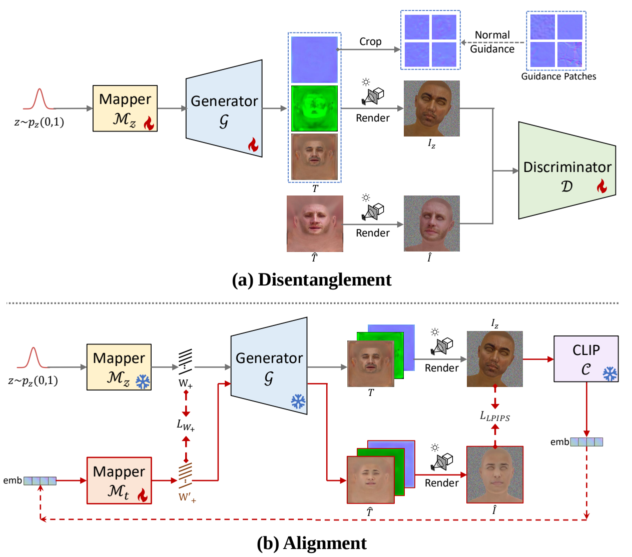


Fig. 4: The disentanglement and alignment pipeline of PBR-GAN. (a) We generate disentangled PBR textures by leveraging entangled FFHQ-UV textures and differentiable rendering. (b) We align the latent space with the text space under the guidance of CLIP to achieve text-guided generation. The red part represents the newly added modules based on (a). The flame represents a trainable module, while the snowflake represents a frozen module.

from entangled FFHQ-UV textures to disentangle PBR textures. (2) Alignment (Sec. III-B). We align the latent space of GAN with the text embedding space using a CLIP-based loss. (3) Boosting (Sec. III-C). We further boost the generated textures with a new edge-aware SDS loss for uncommon text conditions.

A. *Disentanglement*

Given an arbitrary 3D facial geometry, we propose a self-supervised learning scheme to generate disentangled UV textures, *e.g.* , the diffuse/specular/normal maps, via leveraging a non-PBR facial texture dataset FFHQ-UV [10]. FFHQ-UV is constructed with even illumination conditions from facial images. The textures are entangled with weak specular highlights. Directly applying them to train the generation model may introduce light spot artifacts. To avoid this issue, we do not utilize FFHQ-UV data as supervision. Instead, we train a disentangling model with differentiable rendering and additional regularization terms.

As diffusion models typically involve iterative noise injection and denoising processes, the changes in the image are subtle at each iteration. Integrating diffusion models with differentiable rendering during the denoising process may lead to slow training. It hinders us from directly training a diffusion model with FFHQ-UV output. Therefore, we construct a PBR texture generator based on StyleGAN2 [13] for the disentanglement. The consistency among albedo, roughness, and normal is implicitly enforced during the carefully designed training process. The following sections will provide detailed introductions to the network architecture, training strategy, and regularization terms.

a) Network structure.: Fig. 4(a) demonstrates the first stage of our PBRGAN framework. Given a randomly initialized latent code $z \in \mathbb{R}^{512}$, a latent mapper $\mathcal{M}_z$ transforms z into an intermediate space $\mathcal{W}_+$ to produce a new latent embedding $w \in \mathbb{R}^{16*512}$. Then w modulates the parameters of $\mathcal{G}$ through adaptive instance normalization (AdaIN) [74]. To enable StyleGAN with physically-based rendering, a generator $\mathcal{G}$ outputs a seven-channel appearance map in the unwrapped UV space. Following the setting of Munkberg et al. [11], the output can be split into $T = (T_a, T_r, N)$, where T_a is the albedo map, T_r are roughness parameters, and N describes tangent space normal perturbations. In this design, the rendering components share the same feature space to ensure inherent alignment. For the geometry, we generate random facial meshes by sampling in the parameter space of the 3DMM-based model HIFI3D++ [75]. It can provide diverse geometric information and help to ensure the adaptability of generated results to different geometries.

b) Differentiable-rendering-based training.: We introduce a differentiable rendering module $\mathcal{R}$ to train the PBRGAN. $\mathcal{R}$ takes the predicted PBR textures, a 3D geometry s, and a random pose p as input and produces a 2D facial image $I_z = \mathcal{R}(T_a, T_r, N, s, p)$. For the ground truth, $\mathcal{R}$ utilizes an entangled texture map $\hat{T}$ in FFHQ-UV to render another image $\hat{I} = \mathcal{R}(\hat{T}, s, p)$. Then PBRGAN can be supervised by the distance between I_z and $\hat{I}$. Besides, I_z and $\hat{I}$ are input to a global discriminator and a patch-based local discriminator for detail enhancement. We employ the discriminator training technique of Diffusion-GAN [76] to improve the diversity of the results. Through differentiable-rendering-based training, the consistency among albedo, roughness, and normal is further reinforced through constraints based on multi-view rendering results, where any misalignment would result in a higher adversarial loss.

c) Regularization terms.: To generate reasonable texture components, we design regularization terms of albedo and normal to provide prior guidance for the material properties.

(1) Patch-level normal constraint. Due to the absence of the ground truth PBR texture data, the generated normal maps tend to be blurry and lack details. To address this issue, we propose a patch-level normal constraint to provide priors, ensuring the generated normal maps align with the distribution of real normal data. Since normal maps mainly focus on local detail features, we utilize small patches from a general normal image dataset, which can cover the facial normal distribution. Specifically, we collect several $4K$ reference normal maps from the public general object dataset MatSynth [77] for detailed guidance. To ignore the global appearance and focus on the local details, such as bumps, grooves, and scratches, we randomly crop reference normal maps into 32×32 patches. We also utilize data augmentation pipeline (e.g., random scaling, rotation, shifts) to expand the effective diversity of normal map representations. Then we introduce a patch-level normal discriminator to align the distributions of the patches cropped from generated normal maps and those in our dataset. Complex facial structures are decomposed into atomic 32×32 primitives during training, allowing the network to learn the detailed features while maintaining the spatial consistency of different patches. To

verify that the patches in the training set cover the facial normal map details, we conducted a retrieval experiment using a real-world facial normal map. Please refer to the supplemental material for more details.

(2) Albedo constraint. As the textures of FFHQ-UV may suffer from highlight artifacts, we design a regularization term for the albedo $\mathcal{J}_a$ to guide the texture disentanglement. $\mathcal{J}_a$ is used to reduce the specular highlights by constraining the albedo of the skin region to be smooth, which can be represented as:

$$\mathcal{J}_a = |\nabla(T_a \odot M_{skin})| \tag{1}$$

where M_{skin} is the mask of the skin region, $\odot$ is elment-wise multiplication, ∇ represents the total variation. $\mathcal{J}_a$ is calculated in both the RGB space and the luminance channel of the HLS color space.

B. Alignment

With the emergence of pre-trained cross-modal models, text prompts can provide a flexible and user-friendly interface for 3D content generation. Existing approaches typically rely on fine-tuning the generative models to establish consistency between the results and the texts [6]. However, it usually requires additional optimization for different objects or prompts. To avoid this issue, we extend StyleGAN to support text-guided PBR texture generation by correlating the text embeddings with the $\mathcal{W}_+$ latent space of PBRGAN.

Recent advances based on GAN inversion techniques and cross-modal models have led to a more feasible solution for aligning text with latent space. GAN-inversion-based image editing algorithms [78]–[80] invert given images back into the latent space of GANs and manipulate them by varying the latent code in different interpretable directions, such as age and expression. In the meantime, the CLIP model has made remarkable progress in establishing cross-modal correlations between image and text space [81]. Therefore, we employ the CLIP model as an anchor point for alignment between text space and $\mathcal{W}_+$ space.

a) The alignment network.: As shown in Fig. 4(b), we integrate an additional text embedding module into the PBRGAN to achieve the alignment. It consists of a CLIP embedding module $\mathcal{E}$ and a text mapper $\mathcal{M}_t$ with a twining structure of the latent mapper $\mathcal{M}_z$ in Sec. III-A. In this stage, the parameters of $\mathcal{G}$ and $\mathcal{E}$ are frozen. $\mathcal{M}_t$ is optimized to map the text embedding space $\mathcal{W}'_+$ to $\mathcal{W}_+$. To train $\mathcal{M}_t$, we random sample a latent z and produce an image $I_z = \mathcal{R}(\mathcal{G}(z), s, p)$. Then I_z is fed into $\mathcal{E}$ as input to get the embedding t_z in the CLIP space. The text mapper $\mathcal{M}_t$ takes t_z as input and transforms it into $w' \in \mathcal{W}'_+$. w' is utilized to generate a facial image, which can be represented as:

$$I_t = \mathcal{R}(\mathcal{G}(\mathcal{M}_t(\mathcal{E}(I_z))), s, p). \tag{2}$$

b) The alignment loss functions.: The space alignment is supervised by a loss $\mathcal{L}_{\mathcal{W}_+}$ in the $\mathcal{W}_+$ space and a loss $\mathcal{L}_{LPIPS}$ in the image space. $\mathcal{L}_{\mathcal{W}_+}$ is the Manhattan distance between $\mathcal{W}'_+$ and $\mathcal{W}_+$. $\mathcal{L}_{LPIPS}$ is a LPIPS loss to constrain the text-embedding-based rendered image I_t to have similar perceptual appearance with I_z. The overall loss function is as follows:

$$\mathcal{L} = \alpha|\mathcal{M}_t(\mathcal{E}(I_t)) - \mathcal{M}_z(z)| + \beta LPIPS(I_z, I_t), \tag{3}$$

where α,β are the weights for each component. After the training, PBRGAN can generate PBR textures for a given text prompt t through a feedforward pass $\mathcal{G}(\mathcal{M}_t(\mathcal{E}(t)))$.

C. Boosting

On the foundation of the former stages, PBRGAN is proficient at generating images from common text prompts resembling the FFHQ distribution. However, it still faces limitations in producing reasonable textures using uncommon prompts, *e.g.* , "a prominent knife scar on the face". To deal with these prompts, existing methods [54], [55] employ a pre-trained diffusion model as an image prior and perform parameters optimization using the SDS loss [6]. The optimization is usually trained from scratch. This may require hundreds of iterations for each input prompt or encounter challenges to converge to an optimal solution. It reduces the efficiency and accuracy of texture generation.

In the third stage of PBRGAN, we introduce an EASDS-based PBR texture boosting algorithm to expand the generation space to produce more diverse textures using uncommon prompts. Specifically, we use results generated by PBRGAN as the initialized optimization target of the SDS loss (Fig. 5 (a)). It can produce better initial positions for the SDS optimization to reduce the computational cost and shorten the optimization path. Then we modify the basic SDS loss into an enhanced edge-aware version (EASDS) by introducing a key facial structure alignment module and a pre-trained ControlNet ϕ [82]. The official ControlNet-HED [1] is utilized for all experiments. The EASDS comprises of four steps (Fig. 5 (b)):

(1) Optimization target preparation. We run an inference procedure of PBRGAN to generate the texture maps T_0 for a given text prompt τ. The target parameter T to be optimized of EASDS is initialized with T_0.

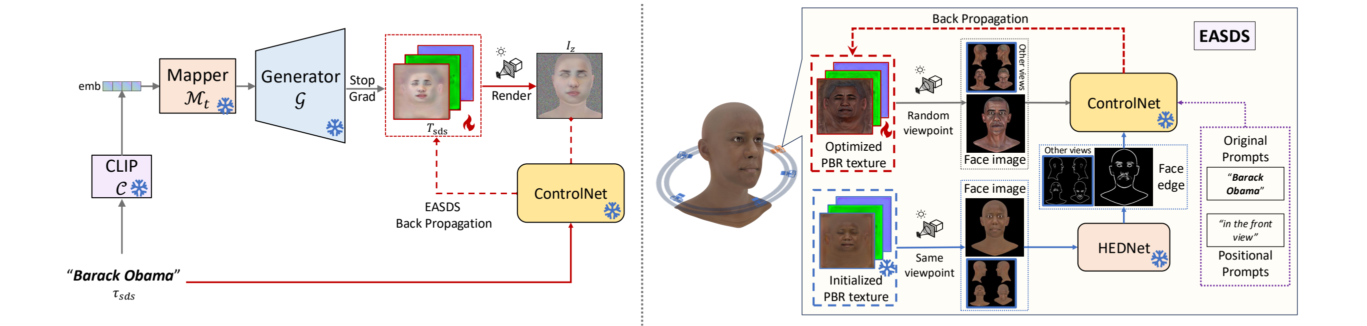
(2) Camera pose setting. We place a series of pre-defined viewpoints around the face model to fit our differentiable-rendering-based pipeline.

(3) Feature-line extraction. We render reference facial images I_{ref} from the viewpoints with a random-sampled geometry and T_0. Then a pre-trained network HEDNet [83] is exploited to detect soft edges E from the rendered images.

(4) EASDS optimization. For a given viewpoint, the rasterizer renders a facial image $I_\tau = \mathcal{R}(T, s, p)$ with the same geometry of (3). The rough HED edges associated with this viewpoint from (3) are taken as feature-line priors to ControlNet. We optimize T with the ControlNet gradient through the rendered image I_τ. The ControlNet ϕ utilizes τ and E to provide a score function $\hat{\epsilon}_\phi$ as the optimization guidance for T:

$$\nabla_T \mathcal{L}_{\text{EASDS}}(\phi, I_\tau) \triangleq \mathbb{E}_{t,\epsilon}\left[w(t)\left(\hat{\epsilon}_\phi\left(\mathbf{z}_t^{I_\tau}; \tau, t, E\right) - \epsilon\right)\frac{\partial I_\tau}{\partial T}\right] \tag{4}$$

where $w(t)$ is a timestep-related weighting function, ϵ is the sample noise, $\mathbf{z}_t^{I_\tau}$ is the noisy image based on I_τ. The optimization starts from the front view and iteratively



(a) EASDS Boosting **(b) EASDS**

Fig. 5: The pipeline of EASDS boosting and details of EASDS. (a) We further boost the generated PBR textures with EASDS. (b) Details of our edge-aware SDS (EASDS). We sample from different viewpoints and employ soft edges detected from a PBRGAN generated texture as feature-line conditions for ControlNet. The flame represents a trainable module, while the snowflake represents a frozen module.

proceeds to other viewpoints. With feature lines extracted from the same 3D reference model, EASDS ensures that the images rendered from different perspectives maintain multi-view consistency in key facial structures.

The pseudocode of the EASDS is shown in Algorithm 1.

Algorithm 1 Edge-aware Score Distillation Sampling

Require: A text prompt τ. A 3D facial geometry s. A trainable PBR texture T initialized with T_0. Camera pose list P. A differentiable renderer $\mathcal{R}$. Learning rate η. The HEDNet $\mathcal{H}$. The HED-based ControlNet ϕ provides a score function $\hat{\epsilon}$.

for p in P **do**
  Extract feature-line $E = \mathcal{H}(\mathcal{R}(T_0, s, p))$
  Render the facial image $I_\tau = \mathcal{R}(T, s, p)$
  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $T \leftarrow T - \eta \mathbb{E}_{t,\epsilon}\left[w(t)\left(\hat{\epsilon}_\phi\left(\mathbf{z}_t^{I_\tau}; \tau, t, E\right) - \epsilon\right)\frac{\partial I_\tau}{\partial T}\right]$
end for
return T

For special text prompts of specific individuals, such as Barack Obama, Avatar, Iron Man, *etc.* , PBRGAN will produce the corresponding texture maps via direct optimization based on EASDS. Furthermore, we could distill relevant knowledge of the EASDS into the latent space of PBRGAN to expand the latent space and achieve more efficient generation. The detailed descriptions can be referred in the supplementary material.

IV. Experiments

In this paper, we focus on text-driven facial PBR texture generation. The experiments are conducted on a Linux server with 8 Tesla V100 GPUs and an Intel Xeon Platinum 8268 CPU. We demonstrate details of experiments, comparisons with state-of-the-art methods, and ablation studies.

A. Experimental Setup

We use the FFHQ-UV dataset [10] for experiments. FFHQ-UV is a large-scale 3DMMs-based facial UV-texture dataset derived from a natural image dataset FFHQ. It contains $50,000+$ texture UV-maps with even illuminations and cleaned facial regions. For the facial geometry, we randomly sample $10\,000$ facial mesh data from HIFI3D++ [75]. Since HIFI3D++ contains no eyeballs, we employ the albedo and geometry of the eyeballs produced from the pipeline of FFHQ-UV [10].

To construct the text space, we define a text prompt dictionary with geometry-independent facial attributes. The dictionary consists of common facial attributes, such as "beard" or "red lips", and uncommon attributes like "scar" or "purple lips". We formulate different text prompts by randomly sampling attributes from the dictionary. For more details, please refer to our supplementary materials.

B. Comparison and Analysis

We compare our method with four state-of-the-art texture generation methods, including Tango [5], Fantasia3D [7], ClipFace [2], and DreamFace [8]. ClipFace can be further divided into two parts: unconditional sampling and conditional optimization, while all the other methods only support conditional generation. Therefore, we first compare our method with ClipFace on the unconditional generation task and then make a comparison with all methods on the text-conditional generation task. The results of DreamFace were obtained from the commercial website https://hyper3d.ai/chatavatar using DreamFaceV1.

a) Metrics.: We use the following metrics from three perspectives to evaluate our method.

- **Non-reference image quality assessment (IQA) metrics:** Non-reference IQA [89] aims to evaluate the quality of images without relying on a reference image. We adopt MUSIQ [90], HyperIQA [85], and CLIPIQA+ [86] for the generated image quality evaluation.
- **Task-specific metrics:** We utilize a tailored metric for face quality measurement, *i.e.* TOPIQ$_{face}$ [87], which is trained with face IQA dataset GFIQA [91], to evaluate the quality of the rendered facial images.
- **Prompt-aware metrics:** We use the CLIPSCORE [88] to measure the alignment between the given prompt and the corresponding generated face.

For these metrics, higher scores indicate better image quality or higher text-image consistency.

b) Unconditional texture generation.: For the text-unrelated generation task, we use non-reference-IQA and task-specific metrics to evaluate our method with $1\,000$ randomly sampled latent code z. As shown in the "without prompt" part of Table I, our method achieves a higher score by 10.3% on TOPIQ$_{face}$, which shows the effectiveness of our PBRGAN. Our method also outperforms ClipFace in terms of the non-reference-IQA metrics, yielding better image quality with enhanced sharpness and intricate details. It demonstrates the effectiveness of our latent space representation.

c) Text-driven texture generation.: For the conditional generation task, our methods were evaluated by randomly sampling 30 prompts from our dictionary to cover different attributes. Specifically, we first synthesize the images with these prompts using all methods. The images are combined with corresponding text prompts to generate text-image pairs. Then, we adopt all the above metrics to evaluate the image quality and text-image consistency of these pairs.

Since PBRGAN is capable of generating PBR texture maps with straightforward inferences, our method strikes a balance between efficiency and performance. As illustrated in the

TABLE I: Qualitative comparison results. We conduct experiments with or without prompts. The best results are highlighted as 1st and 2nd. Our method wins first place in the human perceptual study and inference time. We also achieve the second-best in other metrics for text-conditional generation. Note that DreamFace needs captured PBR data to train and spends five times as much inference time as ours. SDFT is a Stable Diffusion 1.5 model specifically fine-tuned with facial images was added [84].

	Method	Rep	Data	Inf Time / min	Metrics				User Study
					hyperIQA [85] ↑	CLIPIQA+ [86] ↑	TOPIQ$_{face}$ [87] ↑	CLIPSCORE [88] ↑	
w/o prompt	ClipFace [2]	Texture	FFHQ	$<$ 0.01 (1x3090)	0.63	0.55	0.58	-	-
	Ours.	PBR	FFHQUV	$<$ 0.01 (1x3090)	0.74	0.57	0.64	-	-
w/ prompt	TANGO [5]	PBR	-	10 (1x 3090)	0.50	0.45	0.35	0.67	1.52
	Fantasia3D [7]	PBR	-	45 (8x 3090)	0.66	0.52	0.57	0.74	2.04
	ClipFace [2]	Texture	FFHQ	90 (1x 3090)	0.63	0.56	0.52	0.77	2.11
	DreamFace [8]	PBR	LightStage	5 (1x A6000)	0.76	0.63	0.77	0.72	2.78
	Ours	PBR	FFHQUV	$<$ 0.01 (1x3090)	0.68	0.57	0.84	0.76	3.47
	Ours-EASDS	PBR	FFHQUV	1 (1x 3090)	0.66	0.54	0.81	0.74	3.36
	Ours-EASDS+SDFT	PBR	FFHQUV	1 (1x 3090)	0.66	0.57	0.81	0.78	-

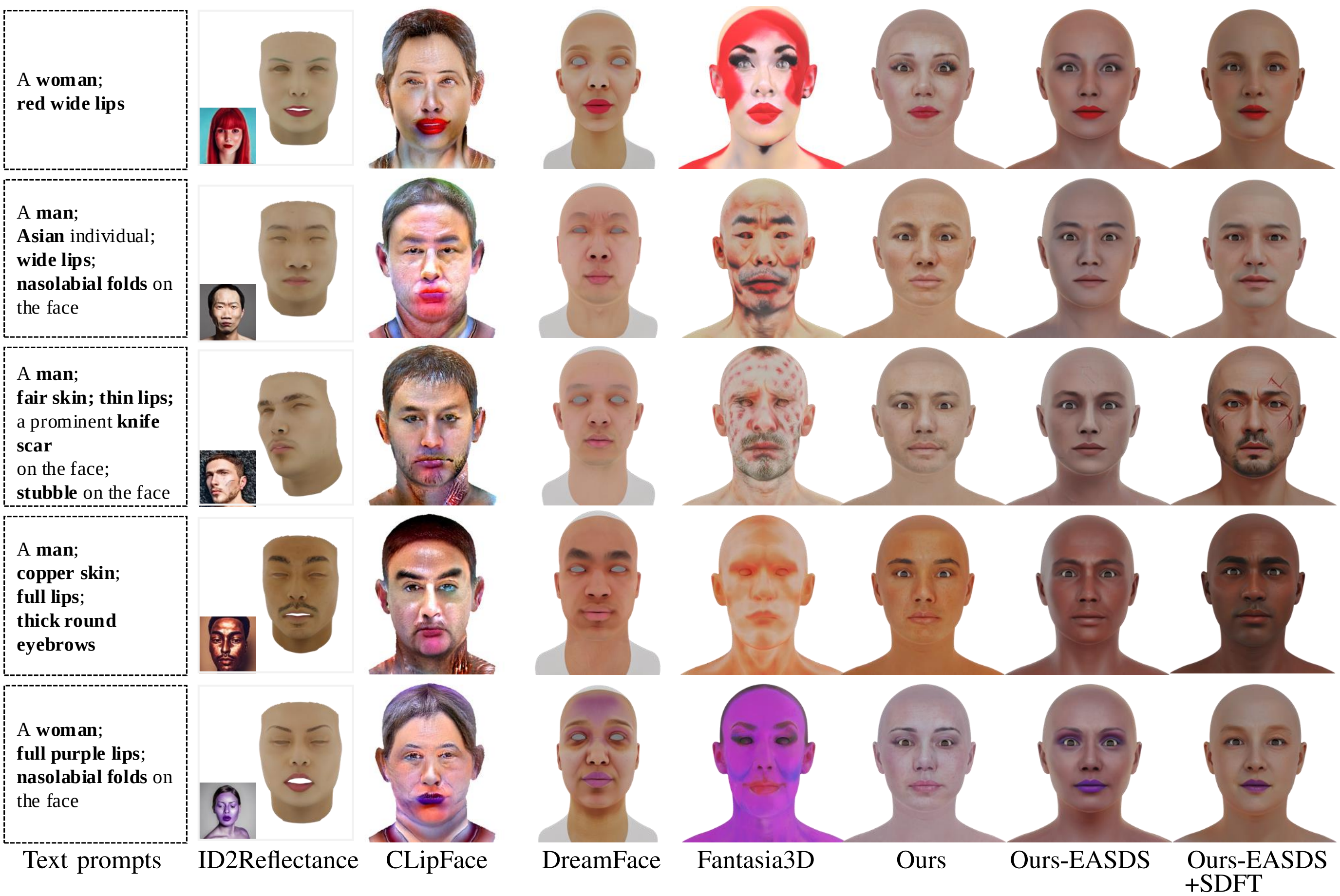


Fig. 6: Qualitative comparison results. From left to right: the input text prompts, the results generated with ClipFace, DreamFace, Fantasia3D, ID2Reflectance, Ours, Ours-EASDS, and Ours-EASDS+SDFT. The results of ID2Reflectance is generated from images produced with Stable Diffusion 1.5.

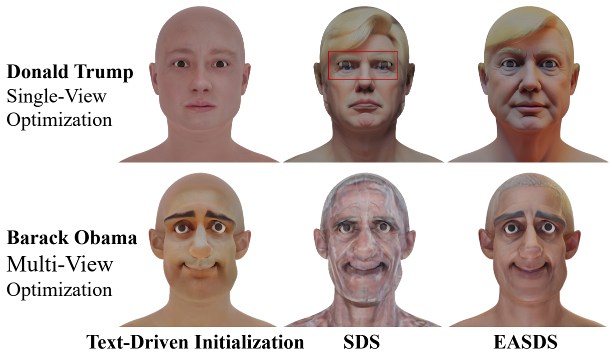


Fig. 7: Ablation study results of EASDS. Note that there are artifacts and misalignments in the results of SDS. The second row shows the generated textures using challenging geometry under the guidance of "Barack Obama". Benefiting from EASDS's alignment of feature lines, our method is capable of generating better textures tailored for extreme geometries.

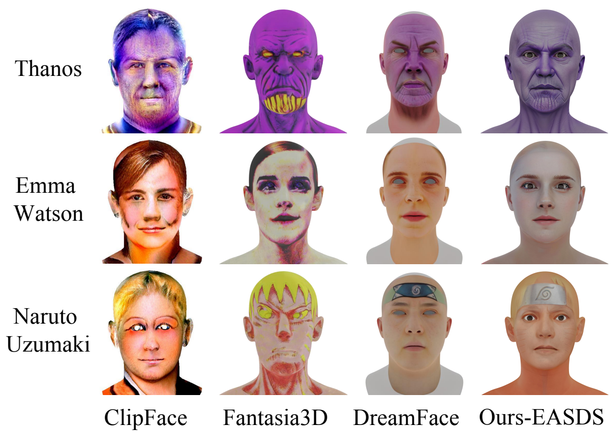


Fig. 8: Comparison of classic celebrity (Emma Watson) and fictional characters (Thanos and Naruto Uzumaki).

"with prompt" part of Tab. I, PBRGAN based on EASDS optimization (Ours-EASDS) is able to generate textures at 1-minute level. It shows that the amalgamation of GAN and SDS can reduce the optimization time by a large margin. For the quality evaluation, Ours-EASDS achieves comparable results to DreamFace with tenuous distinction and outperforms all other methods. Note that DreamFace is trained with captured PBR textures, while our method involves no PBR data. Further expanding the latent space with EASDS-based textures (Ours) may slightly diminish partial quantitative metrics, but it decreases the inference time to less than 0.5 seconds. Fig. 6 demonstrates the qualitative comparison results. Although ClipFace can generate facial textures with most designated attributes, its results have obvious artifacts (the 2nd column). Fantasia3D is designed for generic-object texture generation. Applying it to facial texture generation may cause disharmonious patterns (the 4th column). DreamFace is able to generate relatively realistic results. However, it fails to handle uncommon attributes like "scar" or "purple lips" (the 3rd column). Our methods are capable of producing reasonable facial texture maps with diverse prompts. The last column shows results using EASDS integrated with a facial-specialized Stable Diffusion 1.5 model (fine-tuned on facial datasets) [84], achieving higher conformity with facial descriptions than baseline EASDS. This gain may stem from the gradient

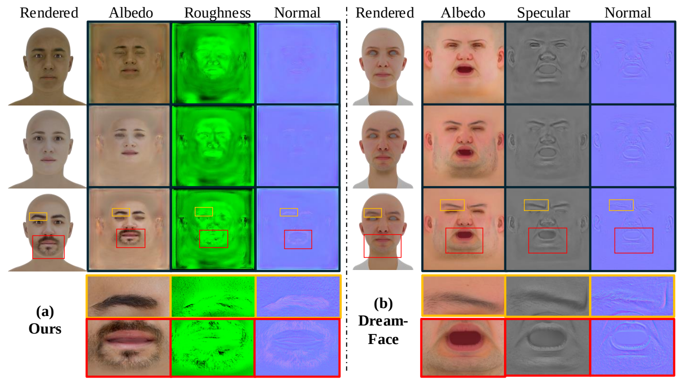


Fig. 9: Disentangled PBR results unconditionally generated with PBRGAN and DreamFace. The results of DreamFace are generated under random seeds with the prompt "Face". Bottom: Magnified views of obvious facial features (beards/eyebrows in red/yellow box) for detailed comparison.

guidance better aligned with descriptions during EASDS optimization. Considering fairness in comparison, we conducted comparative experiments in this paper using EASDS based on the original Stable Diffusion 1.5 model. Furthermore, we also compare our method with text-image-PBR pipeline. We first generate facial images via Stable Diffusion 1.5 based on textual prompts, then estimate PBR textures using an image-to-3D method ID2Reflectance [73]. The experimental results are shown in the second column of Fig. 6. Our method demonstrates superior performance in the textural details, such as purple lips, knife scars, and nasolabial folds. There are two critical limitations in the text-image-PBR pipeline: 1) The subsequent workflow cannot compensate for missing facial features in SD-generated images; 2) While ID2Reflectance effectively preserves fundamental facial features from SD outputs (e.g., eyebrow shape and lip contours), it struggles to transfer unusual characteristics (e.g., scars or makeup colors).

Fig. 8 presents a comprehensive comparison in text-based well-known celebrities generation, highlighting the generalization capabilities of EASDS. The results show that ClipFace and Fantasia3D can generate semantically correct results, but they often have a cartoonish feel. DreamFace performs well in generating results related to real individuals but struggles with the intricate details of fictional characters, such as Thanos's oversmooth appearance and Naruto's missing characteristic whisker marks. Our method, on the other hand, consistently delivers highly realistic and finely detailed results across both real-world celebrity and fictional character generation.

d) Disentangled PBR Results: Fig. 9 exhibited more disentangled PBR results generated by our PBRGAN and DreamFace. The two methods use different geometry unwrapping and rendering representations: Ours employs albedo maps, roughness parameters, and normal maps, while DreamFace uses albedo maps, specular maps, and normal maps, making direct comparisons of these maps difficult. We thus magnify facial features to evaluate PBR map alignment and physical plausibility. Furthermore, while maintaining alignment, our normal maps exhibit plausible geometric variations, and roughness parameters demonstrate logical material properties: beard/eyebrow areas show higher roughness with brighter tones,

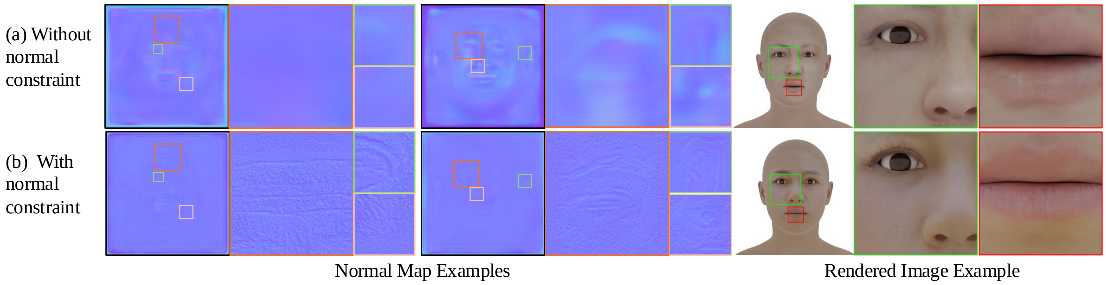


Fig. 10: Ablation study results of normal constraint. The examples are randomly sampled by PBRGAN trained with and without normal constraint separately.

Fig. 11: Ablation Study of our albedo constraint. (a): The results of PBRGAN trained without albedo constraint. (b): The results of PBRGAN trained with albedo constraint.

whereas oil-prone nasal tips/wings display lower roughness with darker coloration.

e) User study.: We conducted a user study to evaluate the perceptual quality of the results. We generate 30 text-image pairs for each method of TANGO, Fantasia3D, ClipFace, DreamFace, Ours, and Ours-EASDS. 100 participants are recruited to evaluate the results. We ask every participant to grade each text-image pair according to the image quality and text-image consistency. The grading is quantified as a preference score to assess the overall quality from 1 (least low quality) to 5 (most high quality). As the last column of Tab. I shows, ours-EASDS is superior to other methods.

C. Ablation Study

We have conducted ablation experiments for disentanglement constraints and EASDS to explore their influence on the quality of the generated PBR textures.

a) Disentanglement Constraints: We conducted ablation study on the patch-level normal constraint and albedo constraint. We compared the generated PBR texture using PBRGAN trained with or without the patch-level normal constraint. The results can be found in Fig. 10, where the three local normal maps on the right side of each sub-image correspond to the marked box in the global image. Local normal maps with normal constraints exhibit sharper texture details compared to those without constraints. These details represent finer facial features such as eyebrows, pores, wrinkles, etc. Rich texture details with the normal constraint (both in the normal maps and rendered images) indicate that the face gains more geometric detail, making it more realistic and closer to real human faces.

To better show the effectiveness of our albedo constraint, we conducted an experiment that trained our PBRGAN with the same setting except for removing the albedo constraint. As shown in Fig. 11, the results of PBRGAN without the albedo constraint have more highlights and overall color inconsistency.

TABLE II: Quantitative results of different constraints. $\mathcal{J}_a/\mathcal{J}_n$ represents albedo/normal constraint, respectively.

Methods	Metrics			
	hyperIQA↑	CLIPIQA+↑	TOPIQ$_{face}$↑	CLIPSCORE↑
Ours w/o $\mathcal{J}_a$ w/o $\mathcal{J}_n$	0.55	0.50	0.82	0.74
Ours w/o $\mathcal{J}_n$	0.59	0.56	0.83	0.74
Ours	0.68	0.57	0.84	0.78
Ours-SDS	0.61	0.54	0.71	0.71
Ours-EASDS	0.66	0.54	0.81	0.74

We also present quantitative comparisons of the ablation study in Tab. II. The metrics further demonstrate the effectiveness of the patch-level normal constraint and albedo constraint.

b) EASDS vs SDS: We compare our EASDS with the classical SDS [6], which has no edge priors. As the Fig. 7 shows, there are some misalignments between the geometry edges and texture edges in the SDS-based results. With additional feature-line priors, our EASDS can enhance facial structural accuracy and generate better PBR textures than SDS, which improves multiple metrics in Tab. II.

D. Limitations and Future Works

PBRGAN leverages the CLIP encoder during the Alignment and EASDS Boosting stages for cross-modal generation. However, due to the CLIP encoder's difficulty in fully embedding long and complex textual information, this reliance leads to some limitations: 1) It struggles to process prompts with more than five attributes effectively. 2) The generated results may not be completely aligned to the text prompts. To mitigate these issues, we plan the following two targeted improvements in the future:

- **Adopting advanced cross-modal embedding:** Replacing CLIP with more advanced and robust models (*e.g.* LongCLIP [92] or TULIP [93]) could enhance the long-text comprehension and alignment accuracy.
- **Facial domain-specific fine-tuning:** For the overlooked facial attributes, we plan to fine-tune the cross-modal alignment network using face-specific datasets such as FFHQ and CelebA with additional rechecked descriptions, to improve local facial feature sensitivity.

V. Conclusion

In this work, we present a multi-stage framework called PBRGAN for text-driven physically based facial texture generation in the wild. We first propose a GAN-based network that can bootstrap from FFHQ-UV textures to generate PBR textures. The network is trained with differentiable rendering and carefully crafted regularization terms to disentangle the FFHQ-UV textures into albedo, roughness, and normal maps without supervision from captured PBR data. Then, we enable PBRGAN with the capability of text-guided generation through latent space and text space alignment. Besides, we introduce an edge-aware SDS loss and an EASDS-based texture boosting scheme to further improve the diversity of the generated results. The experiments demonstrate that our method can efficiently create high-quality and diverse facial PBR textures. In the future, we will investigate the joint generation of precise facial geometries along with diverse PBR textures.

ACKNOWLEDGMENTS

We sincerely thank the anonymous reviewers for their professional and constructive comments, which have helped us improve the quality and clarity of this paper. Weiwei Xu is partially supported by NSFC (No. 62421003) and Ningbo Natural Science Foundation (No. 2025Z062). Rong Zhang is partially supported by NSFC (No. 62302449), Zhejiang Provincial Natural Science Foundation of China (No. LQ23F020009), and "Pioneer" and "Leading Goose" R&D Program of Zhejiang Province (No. 2024C01167). This paper is supported by the Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

REFERENCES

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 10 684–10 695.

[2] S. Aneja, J. Thies, A. Dai, and M. Nießner, "Clipface: Text-guided editing of textured 3d morphable models," in *ACM SIGGRAPH Conf.*, 2023, pp. 1–11.

[3] F. Hong, M. Zhang, L. Pan, Z. Cai, L. Yang, and Z. Liu, "AvatarCLIP: zero-shot text-driven generation and animation of 3D avatars," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 1–19, Jul. 2022.

[4] N. Mohammad Khalid, T. Xie, E. Belilovsky, and T. Popa, "Clip-mesh: Generating textured meshes from text using pretrained image-text models," in *SIGGRAPH Asia Conf.*, 2022, pp. 1–8.

[5] J. Lei, Y. Zhang, K. Jia *et al.*, "Tango: Text-driven photorealistic and robust 3d stylization via lighting decomposition," *Adv. Neural Inform. Process. Syst.*, vol. 35, pp. 30 923–30 936, 2022.

[6] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," in *Int. Conf. Learn. Represent.*, 2023.

[7] R. Chen, Y. Chen, N. Jiao, and K. Jia, "Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation," in *Int. Conf. Comput. Vis.*, 2023, pp. 22 246–22 256.

[8] L. Zhang, Q. Qiu, H. Lin, Q. Zhang, C. Shi, W. Yang, Y. Shi, S. Yang, L. Xu, and J. Yu, "Dreamface: Progressive generation of animatable 3d faces under text guidance," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–16, 2023.

[9] W.-C. Ma, T. Hawkins, P. Peers, C.-F. Chabert, M. Weiss, P. E. Debevec *et al.*, "Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination." *Rendering Techniques*, vol. 2007, no. 9, p. 10, 2007.

[10] H. Bai, D. Kang, H. Zhang, J. Pan, and L. Bao, "Ffhq-uv: Normalized facial uv-texture dataset for 3d face reconstruction," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 362–371.

[11] J. Munkberg, J. Hasselgren, T. Shen, J. Gao, W. Chen, A. Evans, T. Müller, and S. Fidler, "Extracting triangular 3d models, materials, and lighting from images," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 8280–8290.

[12] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 4401–4410.

[13] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 8110–8119.

[14] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," *Adv. Neural Inform. Process. Syst.*, vol. 34, pp. 852–863, 2021.

[15] A. Melnik, M. Miasayedzenkau, D. Makaravets, D. Pirshtuk, E. Akbulut, D. Holzmann, T. Renusch, G. Reichert, and H. Ritter, "Face generation and editing with stylegan: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 3557–3576, 2024.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[17] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 12 873–12 883.

[18] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Adv. Neural Inform. Process. Syst.*, vol. 33, pp. 6840–6851, 2020.

[19] J. Cao, Y. Guo, Q. Wu, C. Shen, J. Huang, and M. Tan, "Improving generative adversarial networks with local coordinate coding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 211–227, 2020.

[20] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Int. Conf. Learn. Represent.*, 2019. [Online]. Available: https://openreview.net/forum?id=B1xsqj09Fm

[21] J.-Y. Zhu, Z. Zhang, C. Zhang, J. Wu, A. Torralba, J. Tenenbaum, and B. Freeman, "Visual object networks: Image generation with disentangled 3d representations," *Adv. Neural Inform. Process. Syst.*, vol. 31, 2018.

[22] Y. Liao, K. Schwarz, L. Mescheder, and A. Geiger, "Towards unsupervised learning of generative models for 3d controllable image synthesis," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 5871–5880.

[23] T. Nguyen-Phuoc, C. Li, L. Theis, C. Richardt, and Y.-L. Yang, "Hologan: Unsupervised learning of 3d representations from natural images," in *Int. Conf. Comput. Vis.*, 2019, pp. 7588–7597.

[24] Y. Shi, D. Aggarwal, and A. K. Jain, "Lifting 2d stylegan for 3d-aware face generation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 6258–6266.

[25] Y. Xu, S. Peng, C. Yang, Y. Shen, and B. Zhou, "3d-aware image synthesis via learning structural and textural representations," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 18 430–18 439.

[26] Y. Deng, J. Yang, J. Xiang, and X. Tong, "Gram: Generative radiance manifolds for 3d-aware image generation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 10 673–10 683.

[27] K. Schwarz, A. Sauer, M. Niemeyer, Y. Liao, and A. Geiger, "Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids," *Adv. Neural Inform. Process. Syst.*, vol. 35, pp. 33 999–34 011, 2022.

[28] E. R. Chan, M. Monteiro, P. Kellnhofer, J. Wu, and G. Wetzstein, "pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 5799–5809.

[29] X. Pan, X. Xu, C. C. Loy, C. Theobalt, and B. Dai, "A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis," *Adv. Neural Inform. Process. Syst.*, vol. 34, pp. 20 002–20 013, 2021.

[30] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis *et al.*, "Efficient geometry-aware 3d generative adversarial networks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 16 123–16 133.

[31] J. Sun, X. Wang, L. Wang, X. Li, Y. Zhang, H. Zhang, and Y. Liu, "Next3d: Generative neural texture rasterization for 3d-aware head avatars," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 20 991–21 002.

[32] Z. Yuan, Y. Zhu, Y. Li, H. Liu, and C. Yuan, "Make encoder great again in 3d gan inversion through geometry and occlusion-aware encoding," in *Int. Conf. Comput. Vis.*, 2023, pp. 2437–2447.

[33] Z. Huang, Q. Chen, L. Sun, Y. Yang, N. Wang, Q. Wu, and M. Tan, "G-nerf: Geometry-enhanced novel view synthesis from single-view images," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 10 117–10 126.

[34] S. Hyun and J.-P. Heo, "Gsgan: Adversarial learning for hierarchical generation of 3d gaussian splats," *Adv. Neural Inform. Process. Syst.*, vol. 37, pp. 67 987–68 012, 2024.

[35] Y. Liu, Z. Shu, Y. Li, Z. Lin, R. Zhang, and S. Kung, "3d-fm gan: Towards 3d-controllable face manipulation," in *Eur. Conf. Comput. Vis.* Springer, 2022, pp. 107–125.

[36] K. Deng, G. Yang, D. Ramanan, and J.-Y. Zhu, "3d-aware conditional image synthesis," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 4434–4445.

[37] F. Tan, S. Fanello, A. Meka, S. Orts-Escolano, D. Tang, R. Pandey, J. Taylor, P. Tan, and Y. Zhang, "Volux-gan: A generative model for 3d face synthesis with hdri relighting," in *ACM SIGGRAPH Conf.*, 2022, pp. 1–9.

[38] B. Deng, Y. Wang, and G. Wetzstein, "Lumigan: Unconditional generation of relightable 3d human faces," in *Int. Conf. on 3D Vis.* IEEE, 2024, pp. 302–312.

[39] K. Sun, S. Wu, N. Zhang, Z. Huang, Q. Wang, and H. Li, "Cgof++: Controllable 3d face synthesis with conditional generative occupancy fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 2, pp. 913–926, 2024.

[40] H. Kim, M. Jang, W. Yoon, J. Lee, D. Na, and S. Woo, "Switchlight: Co-design of physics-driven architecture and pre-training framework for human portrait relighting," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 25 096–25 106.

[41] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou, "Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 1155–1164.

[42] M. Oechsle, L. Mescheder, M. Niemeyer, T. Strauss, and A. Geiger, "Texture fields: Learning texture representations in function space," in *Int. Conf. Comput. Vis.*, 2019, pp. 4531–4540.

[43] K. Schwarz, Y. Liao, M. Niemeyer, and A. Geiger, "Graf: Generative radiance fields for 3d-aware image synthesis," *Adv. Neural Inform. Process. Syst.*, vol. 33, pp. 20 154–20 166, 2020.

[44] L. Gao, T. Wu, Y.-J. Yuan, M.-X. Lin, Y.-K. Lai, and H. Zhang, "Tm-net: Deep generative networks for textured meshes," *ACM Trans. Graph.*, vol. 40, no. 6, pp. 1–15, 2021.

[45] T. Shen, J. Gao, K. Yin, M.-Y. Liu, and S. Fidler, "Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis," *Adv. Neural Inform. Process. Syst.*, vol. 34, pp. 6087–6101, 2021.

[46] J. Gao, T. Shen, Z. Wang, W. Chen, K. Yin, D. Li, O. Litany, Z. Gojcic, and S. Fidler, "Get3d: A generative model of high quality 3d textured shapes learned from images," *Adv. Neural Inform. Process. Syst.*, vol. 35, pp. 31 841–31 854, 2022.

[47] Y. Siddiqui, J. Thies, F. Ma, Q. Shan, M. Nießner, and A. Dai, "Texturify: Generating textures on 3d shape surfaces," in *Eur. Conf. Comput. Vis.* Springer, 2022, pp. 72–88.

[48] A. Lattas, S. Moschoglou, S. Ploumpis, B. Gecer, J. Deng, and S. Zafeiriou, "Fitme: Deep photorealistic 3d morphable model avatars," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 8629–8640.

[49] A. Sanghi, H. Chu, J. G. Lambourne, Y. Wang, C.-Y. Cheng, M. Fumero, and K. R. Malekshan, "Clip-forge: Towards zero-shot text-to-shape generation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 18 603–18 613.

[50] A. Jain, B. Mildenhall, J. T. Barron, P. Abbeel, and B. Poole, "Zero-shot text-guided object generation with dream fields," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 867–876.

[51] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[52] O. Michel, R. Bar-On, R. Liu, S. Benaim, and R. Hanocka, "Text2mesh: Text-driven neural stylization for meshes," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 13 492–13 502.

[53] C. Wang, M. Chai, M. He, D. Chen, and J. Liao, "Clip-nerf: Text-and-image driven manipulation of neural radiance fields," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 3835–3844.

[54] H. Wang, X. Du, J. Li, R. A. Yeh, and G. Shakhnarovich, "Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 12 619–12 629.

[55] G. Metzer, E. Richardson, O. Patashnik, R. Giryes, and D. Cohen-Or, "Latent-nerf for shape-guided generation of 3d shapes and textures," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 12 663–12 673.

[56] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin, "Magic3d: High-resolution text-to-3d content creation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 300–309.

[57] J. Tang, J. Ren, H. Zhou, Z. Liu, and G. Zeng, "Dreamgaussian: Generative gaussian splatting for efficient 3d content creation," in *The Twelfth International Conference on Learning Representations*, 2024.

[58] T. Yi, J. Fang, J. Wang, G. Wu, L. Xie, X. Zhang, W. Liu, Q. Tian, and X. Wang, "Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 6796–6807.

[59] Z. Wang, C. Lu, Y. Wang, F. Bao, C. Li, H. Su, and J. Zhu, "Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation," in *Adv. Neural Inform. Process. Syst.*, 2023.

[60] X. Han, Y. Cao, K. Han, X. Zhu, J. Deng, Y.-Z. Song, T. Xiang, and K.-Y. K. Wong, "Headsculpt: Crafting 3d head avatars with text," *Adv. Neural Inform. Process. Syst.*, vol. 36, pp. 4915–4936, 2023.

[61] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 18 392–18 402.

[62] Y. Wu, Y. Meng, Z. Hu, L. Li, H. Wu, K. Zhou, W. Xu, and X. Yu, "Text-guided 3d face synthesis-from generation to editing," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 1260–1269.

[63] E. Richardson, G. Metzer, Y. Alaluf, R. Giryes, and D. Cohen-Or, "Texture: Text-guided texturing of 3d shapes," in *ACM SIGGRAPH Conf.*, 2023, pp. 1–11.

[64] S. Yamaguchi, S. Saito, K. Nagano, Y. Zhao, W. Chen, K. Olszewski, S. Morishima, and H. Li, "High-fidelity facial reflectance and geometry inference from an unconstrained image," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–14, 2018.

[65] R. Li, K. Bladin, Y. Zhao, C. Chinara, O. Ingraham, P. Xiang, X. Ren, P. Prasad, B. Kishore, J. Xing *et al.*, "Learning formation of physically-based face attributes," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 3410–3419.

[66] A. Chen, Z. Chen, G. Zhang, K. Mitchell, and J. Yu, "Photo-realistic facial details synthesis from single image," in *Int. Conf. Comput. Vis.*, 2019, pp. 9429–9439.

[67] A. Lattas, S. Moschoglou, S. Ploumpis, B. Gecer, A. Ghosh, and S. Zafeiriou, "Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9269–9284, 2021.

[68] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. ACM, 2023, pp. 157–164.

[69] P. Garrido, M. Zollhöfer, D. Casas, L. Valgaerts, K. Varanasi, P. Pérez, and C. Theobalt, "Reconstruction of personalized 3d face rigs from monocular video," *ACM Trans. Graph.*, vol. 35, no. 3, pp. 1–15, 2016.

[70] L. Tran and X. Liu, "Nonlinear 3d face morphable model," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 7346–7355.

[71] F. P. Papantoniou, A. Lattas, S. Moschoglou, and S. Zafeiriou, "Relightify: Relightable 3d faces from a single image via diffusion models," in *Int. Conf. Comput. Vis.*, 2023, pp. 8806–8817.

[72] A. Dib, L. G. Hafemann, E. Got, T. Anderson, A. Fadaeinejad, R. M. Cruz, and M.-A. Carbonneau, "Mosar: Monocular semi-supervised model for avatar reconstruction using differentiable shading," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 1770–1780.

[73] X. Ren, J. Deng, Y. Cheng, J. Guo, C. Ma, Y. Yan, W. Zhu, and X. Yang, "Monocular identity-conditioned facial reflectance reconstruction," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 885–895.

[74] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Int. Conf. Comput. Vis.*, 2017, pp. 1501–1510.

[75] Z. Chai, H. Zhang, J. Ren, D. Kang, Z. Xu, X. Zhe, C. Yuan, and L. Bao, "Realy: Rethinking the evaluation of 3d face reconstruction," in *Eur. Conf. Comput. Vis.* Springer, 2022, pp. 74–92.

[76] Z. Wang, H. Zheng, P. He, W. Chen, and M. Zhou, "Diffusion-GAN: Training GANs with diffusion," in *The Eleventh International Conference on Learning Representations*, 2023.

[77] G. Vecchio and V. Deschaintre, "Matsynth: A modern pbr materials dataset," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024.

[78] J. Zhu, Y. Shen, D. Zhao, and B. Zhou, "In-domain gan inversion for real image editing," in *Eur. Conf. Comput. Vis.* Springer, 2020, pp. 592–608.

[79] R. Abdal, Y. Qin, and P. Wonka, "Image2stylegan++: How to edit the embedded images?" in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 8296–8305.

[80] Y. Shen and B. Zhou, "Closed-form factorization of latent semantics in gans," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 1532–1540.

[81] J. N. Pinkney and C. Li, "clip2latent: Text driven sampling of a pre-trained stylegan using denoising diffusion and clip," in *Brit. Mach. Vis. Conf.* BMVA Press, 2022, p. 594.

[82] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Int. Conf. Comput. Vis.*, 2023, pp. 3836–3847.

[83] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Int. Conf. Comput. Vis.*, 2015, pp. 1395–1403.

[84] Merjic, "majicmix realistic v7," [Online]. Available: https://civitai.com/models/43331/majicmix-realistic, 2025.

[85] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2020.

[86] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *AAAI*, 2023.

[87] C. Chen, J. Mo, J. Hou, H. Wu, L. Liao, W. Sun, Q. Yan, and W. Lin, "Topiq: A top-down approach from semantics to distortions for image quality assessment," *IEEE Trans. Image Process.*, 2024.

[88] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "CLIPScore: a reference-free evaluation metric for image captioning," in *EMNLP*, 2021.

[89] C. Chen and J. Mo, "IQA-PyTorch: Pytorch toolbox for image quality assessment," [Online]. Available: https://github.com/chaofengc/IQA-PyTorch, 2022.

[90] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "Musiq: Multi-scale image quality transformer," in *Int. Conf. Comput. Vis.*, 2021, pp. 5148–5157.

[91] S. Su, H. Lin, V. Hosu, O. Wiedemann, J. Sun, Y. Zhu, H. Liu, Y. Zhang, and D. Saupe, "Going the extra mile in face image quality assessment: A novel database and model," *IEEE Trans. Multimedia*, 2023.

[92] B. Zhang, P. Zhang, X. Dong, Y. Zang, and J. Wang, "Long-clip: Unlocking the long-text capability of clip," in *Eur. Conf. Comput. Vis.* Springer, 2024, pp. 310–325.

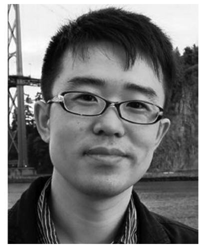
[93] Z. Tang, L. Lian, S. Eisape, X. Wang, R. Herzig, A. Yala, A. Suhr, T. Darrell, and D. M. Chan, "Tulip: Towards unified language-image pretraining," *arXiv preprint arXiv:2503.15485*, 2025.

Chi Wang received a Ph.D. Degree at the state key lab of CAD&CG in Zhejiang University in 2024. He received an M.S. degree in Computer Science from Beijing Institute of Technology in 2017, and a B.S. degree in Software Engineering from Beijing Institute of Technology in 2013. His research interests include 3D asserts, scene, and video generation.

Junming Huang received a Bachelor of Engineering degree from the School of Software, Yunnan University in 2023. He is currently pursuing a Master's degree at the State Key Lab of CAD&CG, Zhejiang University, China. His research interests include computer vision and generative models.

Rong Zhang is currently an associate researcher at Zhejiang Gongshang University. She received a Ph.D. Degree at state key lab of CAD&CG in Zhejiang University in 2022, and B.S. degree in Software Engineering from Tianjin University in 2013. Her research interests include image generation and 3D generation.

Qi Wang received a Bachelor's and Master of Engineering degree from the School of Software Engineering, Beijing Institute of Technology, China in 2017. He is currently a Ph.D. degree candidate in computer graphics at the State Key Lab of CAD&CG, Zhejiang University, China. His research interest is realistic rendering.



Pengfei Wan is the head of the Visual Generation and Interaction Center of Kuaishou Technology. He received a B.E. degree in electronic engineering and information science from the University of Science and Technology of China, Hefei, China, and a Ph.D. degree in electronic and computer engineering from the Hong Kong University of Science and Technology, Hong Kong. His research interests include image/video signal processing, computational photography, and computer vision.

Haotian Yang received the M.S. degree and B.S. degree from the School of Electronic Science and Engineering, Nanjing University, China, in 2021 and 2018 respectively. His research interests include computer vision and computer graphics. He is now a research engineer in Kuaishou Technology.

Chongyang Ma received a B.S. degree from the Fundamental Science Class (Mathematics and Physics) of Tsinghua University in 2007 and a Ph.D. degree in Computer Science from the Institute for Advanced Study of Tsinghua University in 2012. He is currently a Research Lead with Kuaishou Technology, Beijing. His research interests include computer graphics and computer vision.

Haibin Huang received his B.S. and M.S. degrees in Mathematics in 2009 and 2011 respectively from Zhejiang University. He obtained his Ph.D. in Computer Science from UMass Amherst in 2017. He is currently a Senior Staff Research Scientist at Kuaishou Technology.

Weiwei Xu is currently a Professor at the state key lab of CAD&CG in Zhejiang University. He was a Qianjiang Professor at Hangzhou Normal University and a researcher in the Internet Graphics Group at Microsoft Research Asia from 2005 to 2012, and he was a post-doc researcher at Ritsmeikan University in Japan for more than one year. He received a Ph.D. Degree in Computer Graphics from Zhejiang University, Hangzhou, and a B.S. Degree and Master Degree in Computer Science from Hohai University in 1996 and 1999 respectively.