AvatarTex: High-Fidelity Facial Texture Reconstruction from Single-Image Stylized Avatars

Yuda Qiu[1*] Zitong Xiao[1*] Yiwei Zuo[1] Zisheng Ye[1] Weikai Chen[†] Xiaoguang Han[1,2‡]
[1]SSE, CUHKSZ [2]FNii, CUHKSZ

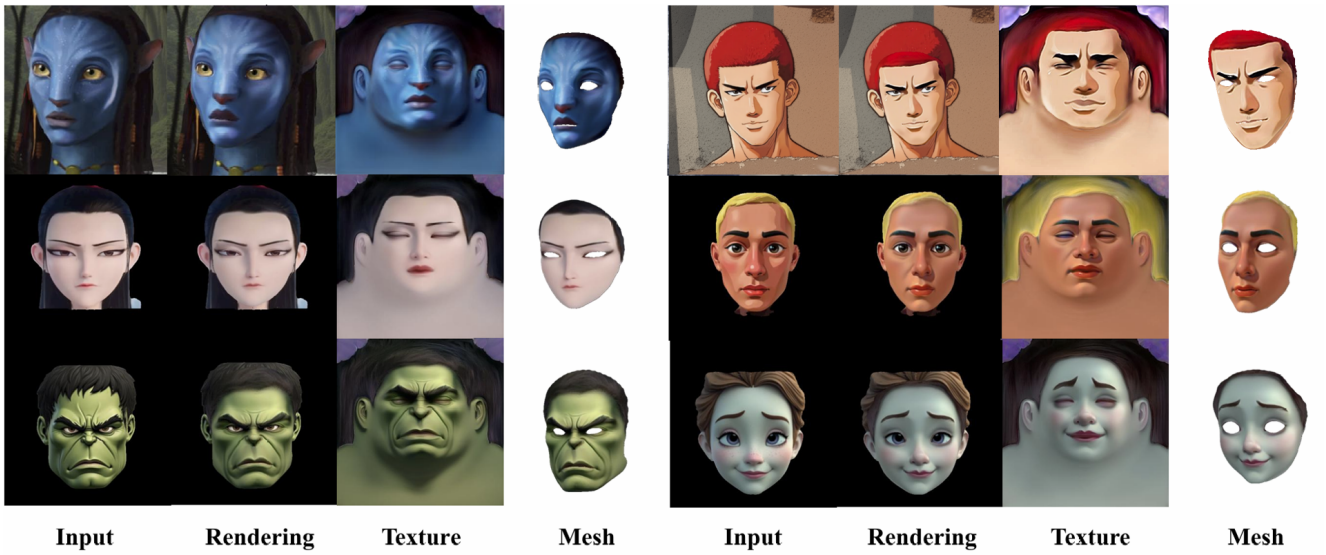

Figure 1. Given an input facial image, AvatarTex generates the corresponding high-fidelity and topology-consistent texture with both artistic and geometric coherence. AvatarTex supports the reconstruction from in-the-wild face images across diverse styles.

Abstract

We present AvatarTex, a high-fidelity facial texture reconstruction framework capable of generating both stylized and photorealistic textures from a single image. Existing methods struggle with stylized avatars due to the lack of diverse multi-style datasets and challenges in maintaining geometric consistency in non-standard textures. To address these limitations, AvatarTex introduces a novel three-stage diffusion-to-GAN pipeline. Our key insight is that while diffusion models excel at generating diversified textures, they lack explicit UV constraints, whereas GANs provide a well-structured latent space that ensures style and topology consistency. By integrating these strengths, AvatarTex achieves high-quality topology-aligned texture synthesis with both artistic and geometric coherence. Specifically, our three-stage pipeline first completes missing texture regions via diffusion-based inpainting, refines style and structure consistency using GAN-based latent optimization, and enhances fine details through diffusion-based repainting. To address the need for a stylized texture dataset, we introduce TexHub, a high-resolution collection of 20,000 multi-style UV textures with precise UV-aligned layouts. By leveraging TexHub and our structured diffusion-to-GAN pipeline, AvatarTex establishes a new state-of-the-art in multi-style facial texture reconstruction. TexHub will be released upon publication to facilitate future research in this field.

1. Introduction

3D facial modeling and texture generation is a fundamental problem in computer vision with widespread applications in gaming, virtual reality, digital humans, and animation. While significant progress has been made in reconstructing realistic 3D facial meshes from monocular images, generating high-fidelity and topology-consistent textures remains a

*Equal Contribution, listed in alphabetical order.

†This paper solely reflects the author's personal research and is not associated with the author's affiliated institution.

‡**Corresponding email:** hanxiaoguang@cuhk.edu.cn

major challenge, especially for stylized facial avatars. Unlike real-world human faces, stylized avatars often feature exaggerated shapes, abstract shading, artistic brush strokes, and varying levels of realism, making traditional texture reconstruction techniques ineffective.

Existing methods, such as FFHQ-UV[1], UV-IDM[21], and UltrAvatar[38], primarily focus on photorealistic texture synthesis and struggle with artistic or stylized representations due to two key limitations: (1) the lack of diverse multi-style texture datasets and (2) the difficulty in handling non-standard texture distributions while maintaining geometric consistency. A straightforward approach for facial texture generation is to directly project RGB color values from an input image onto a 3D mesh, but this leads to self-occlusion artifacts and incomplete texture maps, requiring additional inpainting. However, traditional inpainting methods, including GAN-based [10, 14] and diffusion-based [7, 8, 30, 31] models, often fail to preserve the original artistic style and introduce texture inconsistencies in the occluded regions. In addition, diffusion models alone lack explicit UV constraints, making it difficult to ensure alignment across different facial mesh topologies.

To overcome these challenges, we propose AvatarTex, a novel high-fidelity texture reconstruction framework that supports both stylized and realistic 3D facial texture generation from a single input image. Unlike existing approaches, which struggle with incomplete texture, style inconsistency, and geometric misalignment, AvatarTex combines diffusion-based texture synthesis with StyleGAN-driven latent optimization to ensure high-quality topology-aligned texture generation. Our key insight is that diffusion models excel at generating diverse textures but lack structure-aware consistency, while StyleGAN's latent space provides a well-regularized texture manifold, but struggles with artistic variations. In particular, AvatarTex employs a structured three-stage framework. First, we extract a partial UV texture map from the input image and complete the missing regions using a diffusion-based inpainting network. This provides an initial, but potentially inconsistent texture estimate. To refine style consistency and correct misalignment, we optimize the texture in StyleGAN2's latent space, leveraging its structured texture manifold to enforce coherent alignment in the UV space. Finally, we enhance high-frequency details using diffusion-based repainting, ensuring that the reconstructed texture maintains both global artistic fidelity and fine-scale realism. By integrating these components, AvatarTex bridges the gap between generative flexibility and structural consistency, producing high-quality and topology-aligned textures.

A key component of our approach is the *TexHub* dataset, the first multi-style facial texture data collection that serves as a foundation for training our diffusion-based inpainting network and StyleGAN-based texture refinement. Unlike existing UV texture datasets that primarily focus on photorealistic human faces, TexHub provides a diverse collection of 20,000 high-resolution, multi-style UV texture maps, covering a wide range of artistic and exaggerated facial textures. The dataset is generated using a LoRA-enhanced FLUX[17] model guided by a Canny-based ControlNet, trained on professionally curated artist-designed texture maps. This enables controlled texture synthesis aligned with pre-defined UV layouts. We will release TexHub to benefit future research in multi-style facial texture generation.

Through the integration of our curated dataset and the specially tailored facial texture generation framework, AvatarTex achieves state-of-the-art performance in both photorealistic and stylized facial texture reconstruction, outperforming prior method in terms of both topology consistency and detail fineness. We summarize our contributions as follows:

- We introduce *TexHub*, the first dataset with 20,000 high-quality facial textures designed specifically for multi-style facial texture synthesis, enabling models to generalize beyond photorealistic textures.
- We provide a comprehensive analysis of the optimization behavior in diffusion and GAN latent spaces, revealing the optimization dilemma in the diffusion latent space for texture reconstruction.
- Based on the insight above, we present AvatarTex, a structured three-stage texture reconstruction pipeline with a novel diffusion-to-GAN pipeline that connects the good ends of diffusion and GAN models to maximize their performance in the domain of facial texture synthesis.
- We set the new state of the art in the task of reconstructing facial textures with diverse artistic styles from a single avatar image.

2. Related Works

Image-to-3D Generation Reconstructing high-quality 3D meshes from single-view images remains a core challenge in computer graphics and vision. Recent progress exploits data-driven 2D diffusion models to bridge the 2D–3D gap. DreamFusion[27] pioneered this direction with Score Distillation Sampling (SDS), distilling geometry and appearance priors from pretrained 2D diffusion models via differentiable rendering, proving that purely 2D supervision can enable 3D generation. Zero123[22] extended Stable Diffusion to novel-view synthesis conditioned on relative camera poses. For greater efficiency and 3D consistency, methods such as SyncDreamer[23], Wonder3D[24], and Unique3D[35] fine-tune 2D diffusion models on large-scale 3D data to produce multi-view consistent images, followed by sparse-view reconstruction. Despite impressive geometry, generating topologically consistent, continuous textures

without multi-face artifacts (Janus problem) remains difficult. In our work, we use Unique3D[35] to obtain reference facial geometry, then deform a template mesh to match the target shape.

Realistic Face Reconstruction In 3D face reconstruction and generation, methods generally fall into two categories: improving the *accuracy* of reconstruction from 2D images, and developing more *efficient and scalable* generation techniques.

For accuracy-oriented methods, a major challenge is precise alignment of facial features under complex conditions such as extreme expressions. Recent works employ facial segmentation to guide reconstruction. For example, Part Re-projection Distance Loss (PRDL)[33] converts segmentation into 2D point sets and optimizes their distribution to match target geometry. Approaches like 3DDFA-v2[29] and DECA[5] minimize 3D errors but often lack pixel-level precision when landmarks are sparse or inaccurate. While these methods perform well in realistic single-expression cases, they rarely support diverse styles or exaggerated geometries.

Efficiency-focused methods, such as GGHead[16], leverage 3D Gaussian Splatting[15] to generate realistic human heads from a single 2D image with high speed and consistency, avoiding the heavy computation of traditional 3D GANs. However, similar to reconstruction approaches, GGHead targets high-fidelity realistic heads and cannot produce multi-style or extreme shapes. DiffPortrait3D[6] enables multi-style novel-view synthesis, yet maintaining view consistency remains challenging.

Stylized Face Reconstruction For stylized face reconstruction, early methods proposed constructing dedicated parametric models tailored to specific non-realistic domains. For instance, Qiu et al.[28] introduced a 3D caricature dataset by manually sculpting approximately 2,000 exaggerated 3D face meshes that mimic the characteristics of 2D caricature illustrations. Building on this dataset, Jung et al.[12] developed a neural parametric model for 3D caricatures, which learns a deformation space capable of representing the highly exaggerated facial geometries common in caricature art.Jang et al. [11] presented an automated framework that can generate full-head 3D Toonify style avatars and support GAN-based 3D facial expression editing. These methods marked an important step toward stylized face reconstruction, demonstrating that facial exaggerations could be encoded into learnable representations.

While both of the realistic and stylized approaches significantly advance 3D face modeling, they share a common limitation: their focus is primarily on realistic or specific style (e.g., caricature or Toonify) representations of human faces and they struggle with generating a wide range of styles. Our work addresses this gap by enabling the generation of diverse, stylized 3D meshes and textures that can handle a broader spectrum of facial geometries and artistic styles, providing a more flexible and creative solution to 3D face modeling.

Facial Texture Reconstruction Accurate facial texture reconstruction from 3D facial geometry remains a critical task for photorealistic avatar creation. Conventional approaches employ projective rendering techniques that directly map mesh vertex colors to 2D texture planes, yet struggle to capture high-frequency details and realistic material properties. The advent of differentiable rendering has enabled data-driven breakthroughs, where self-supervised frameworks (e.g., Deep3D[3], DECA[5]) jointly optimize 3D Morphable Model (3DMM) parameters and neural textures by comparing rendered outputs with input images. These methods demonstrate remarkable generalization by leveraging statistical texture priors learned from in-the-wild facial images, though their reconstruction fidelity remains bounded by the expressiveness of linear texture bases.

To overcome the limitations of parametric models, recent works explore non-linear texture representations through adversarial learning. FitMe[19] pioneers this direction by constructing a GAN-based morphable texture space that decouples identity and illumination attributes, achieving enhanced detail synthesis. Parallel efforts adopt refinement-based pipelines: Initial texture estimates from statistical models are progressively enhanced using neural networks - NextFace[4] and HRN[20] employ CNN-based refiners supervised by photometric losses, while AvatarMe++[18] introduces adversarial training with high-quality UV texture datasets. The state-of-the-art UV-IDM[21] further integrates diffusion models to jointly inpaint missing regions and enhance texture resolution through iterative denoising.

Current methodologies reveal a critical dependency on high-quality texture datasets. FFHQ-UV[1] addresses this by synthesizing large-scale photorealistic UV maps through StyleGAN2-driven[13] multiview fusion, establishing a benchmark for texture learning.

Existing approaches primarily focus on photorealistic texture synthesis, often struggling to generate stylized or artistic facial textures that deviate from real-world appearances. This limitation arises partly due to their reliance on domain-specific datasets (e.g., BFM-UV for UV-IDM[21]) or geometric constraints (e.g., FLAME[5] in VGG-Tex[34]), which limit stylistic diversity. In contrast, our work proposes a high-quality multi-style UV texture dataset. We further design a framework to bridge diffusion models' generative flexibility with StyleGAN[13]'s controllability, enabling high-fidelity synthesis of both realistic and stylized textures while maintaining geometric coherence.

3. Methodology

Our AvatarTex aims to reconstruct the UV-texture map from a single in-the-wild image for both stylized and realistic faces. We describe the method in the following two sections: building a novel multi-style facial UV-texture dataset TexHub (Sec. 3.1) and extracting target UV-texture from the target image (Sec. 3.2). Given an input stylized portrait I, our method reconstructs a 3D mesh consistent with the topology M_v and synthesizes a high-fidelity UV texture map T_v using a structured three-stage approach. We show the illustration of TexHub and AvatarTex in Fig. 2.

3.1. TexHub - Dataset for Multi-Style Facial UV-Texture synthesis

Latent Diffusion Models (LDMs) have demonstrated great generalization capabilities and detail generation in diverse image synthesis tasks. In facial texture generation, recent works like UltraAvatar[38] and UV-IDM[21] have successfully adapted LDM for photorealistic facial texture synthesis. However, these approaches require extensive training datasets of facial textures. For multi-style facial texture, currently there is no open-source datasets with sufficient diversity to support the training of LDM. Consequently, acquiring multi-style texture data becomes critical for this task.

Inspired by LoRA[9] and ControlNet[37], we establish a multi-style facial texture generation workflow. We first employ professional artists to create texture maps adhering to the facial topology and UV layout of HiFi3D++[2], as shown in Fig. 3a. These texture maps are used to train a LoRA model that guides a diffusion model FLUX[17] to generate textures spatially aligned with the target UV layout of HiFi3D++[2]. To further maintain the UV layout in the periocular, lip, and nasal regions, we integrate a canny-based ControlNet to guide the generation of UV texture. Specifically, we train our LoRA for 5,280 steps, setting the LoRA strength to 0.8 and the ControlNet to 0.5. During the sampling process, we use the Euler scheduler and additional text to control the style of generated texture.

As shown in Fig. 3b, this framework produces diverse multi-style textures while preserving the UV structure of HiFi3D++. Notably, our method requires only 80 manually crafted textures to generate style-agnostic textures via text prompts without compromising the native generative capabilities of FLUX[17]. Comparative tests confirm that equivalent workflows using SD 1.5 or SD 2.1 fail to achieve comparable quality. We synthesize 20,000 texture maps in 1024x resolution through this pipeline, forming our dataset *TexHub* for subsequent experiments. The details of the text conditions could be found in Sec. 1 of our supplementary.

3.2. AvatarTex - Generalizable Facial Texture Reconstruction

To extract the facial texture faithfully, we first recover the face geometry from a target image I. Recent advancements in multi-view diffusion models have demonstrated remarkable progress in 3D generation tasks, as exemplified by Unique3D[35]. Given a single input image, Unique3D generates highly faithful 3D models with particularly impressive performance in facial region reconstruction. Building upon Unique3D, we develop a topology-consistent facial geometry alignment framework. For an input facial image I, we first utilize Unique3D to obtain an initialized mesh model M. We perform non-rigid iterative closest point (NICP) optimization to deform a template mesh with target topology into the geometric configuration of M. This pipeline guarantees that the generated mesh M_v preserves the target topological structure while adapting to the geometric details of input I, achieving consistent topology across arbitrary input images. Note that our template is adapted from HiFi3d++[2], facilitating seamless integration of our generated texture data with existing high-quality photorealistic facial skin textures.

Combining a UV-texture dataset TexHub and the recovered geometry M_v, a straightforward approach involves inpainting incomplete textures obtained via geometric projection sampling. Although inpainting is usually regarded as a low-level vision task, which could leverage local information, our experiments reveal poor generalization in both GAN (Pix2PixHD)[32] and diffusion (SD 2.1)[30] models trained on TexHub. This indicates the inherent data demands of this inpainting exceed current model capacities. Although diffusion-based inpainting produces globally coherent textures, it lacks fine-grained detail fidelity. To address this, we propose an optimization framework, AvatarTex, which emphasizes constraints in target images. Our final texture reconstruction pipeline comprises the following three stages: texture initialization, texture correction and texture enhancement, specially designed based on the complementary strengths of diffusion model and StyleGAN2.

Our three-stage pipeline is specifically designed to leverage the complementary strengths of diffusion models and StyleGAN2. The texture initialization stage uses a diffusion model's ability to plausibly fill large, missing UV regions from partial projections. However, diffusion models lack a structured latent space, which can lead to style inconsistencies and misalignments. The texture correction stage addresses this by operating in the semantically meaningful latent space of StyleGAN2, allowing for precise refinement of style and geometric alignment. Yet, StyleGAN2 tends to suppress high-frequency details, especially when trained on limited data. The comparison of the process of optimization is shown in Fig. 4. Therefore, the final texture enhancement stage reintroduces these fine-grained details us-

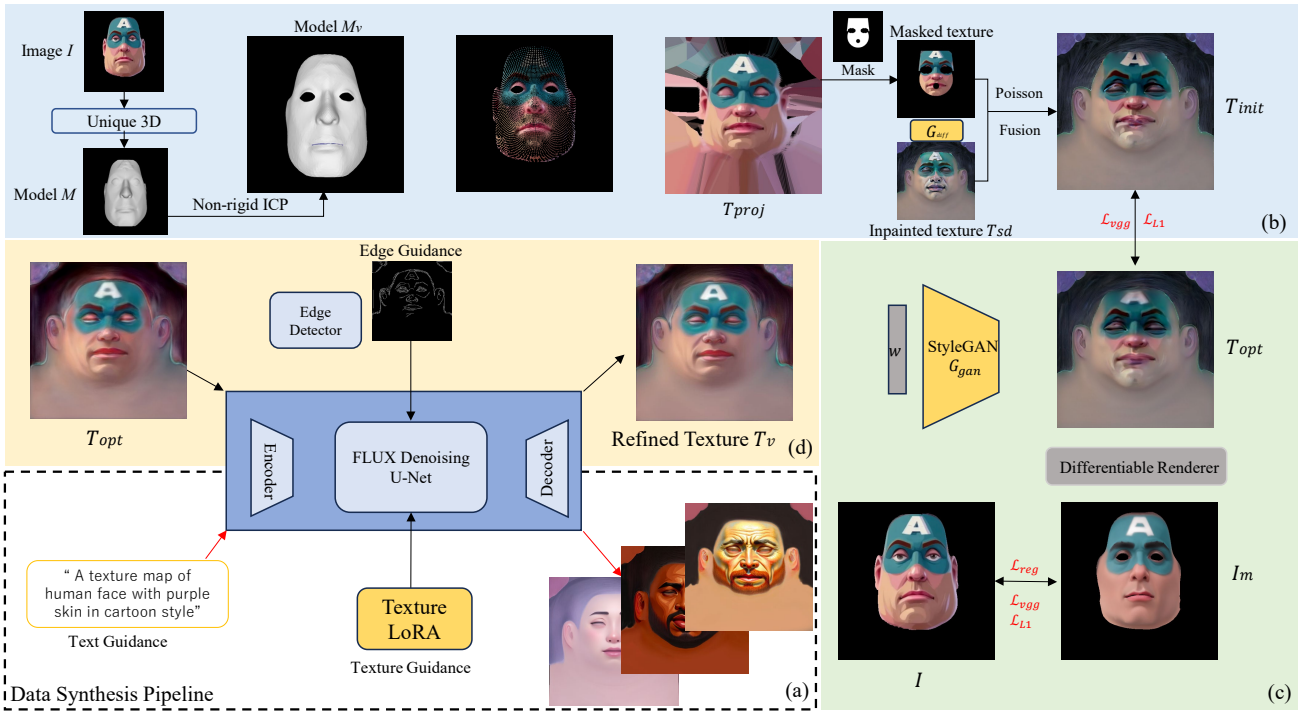

Figure 2. The illustration of our (a) TexHub and (b, c, d) AvatarTex, including (b) texture initialization (c) texture correction and (d) texture enhancement.

(a) Hand Crafted Texture

(b) Generated Texture

Figure 3. The visualization of our TexHub. We guide the FLUX diffusion to generate UV facial texture with LoRA trained with limited hand crafted texture data.

ing SDEdit-based diffusion[25], which sharpens the texture while preserving the corrected structure from the previous stage. This sequential design, based on the unique characteristics of each model, ensures a high-quality, high-fidelity final output.

We illustrate the details of each stage in the following subsections.

3.2.1 Texture Initialization

We train a texture inpainting model G_{diff} based on TexHub. Leveraging FFHQ-UV fitting process, we generate a comprehensive set of masking patterns. During training, a texture map x is randomly sampled from TexHub and processed with a randomly selected mask to create its corrupted counterpart x'. Both x and x' are encoded into latent representations y and y' through the Variational AutoEncoder (VAE) of SD 2.1. We train the UNet part of SD 2.1 to learn the mapping from y' to y, effectively training the model to recover complete textures from partial observations.

During the testing process, given a target image I and its associated mesh M_v, we first project M_v onto the image plane of I to establish 2D-3D correspondence, obtaining positional coordinates P_v. RGB values at these coordinates are then sampled to construct the incomplete texture map T_{proj}. This partial texture is fed into G_{diff} to generate the completed texture T_{sd}. To preserve fine-grained details from the target image, we apply Poisson Image fusion[26] between T_{proj} and T_{sd}, resulting in the initialized texture T_{init}, which optimally combines global structural coherence with local photometric fidelity.

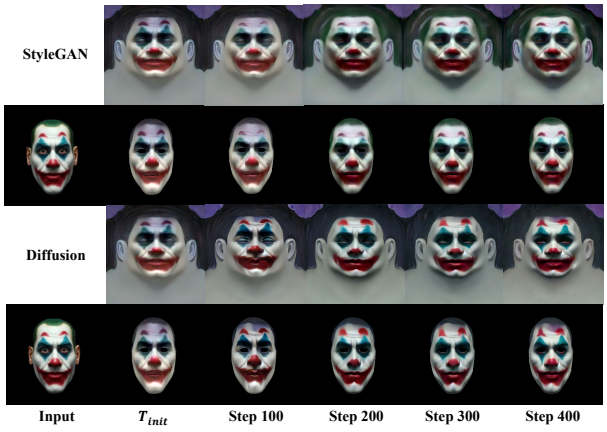

Figure 4. The visualization of the optimization process. The optimization based on diffusion backbone struggles to capture the accurate shape of the local structure, like brows and lips. Instead, the results based on StyleGAN backbone reconstruct the correct features but fail to achieve high quality. The StyleGAN case corresponds to the optimization process in Fig. 6b, and the Diffusion case corresponds to that in Fig. 6e.

3.2.2 Texture Refinement and Correction

While T_{init} provides texture approximations for non-visible regions that are broadly consistent with visible areas, two critical limitations persist:

- The stochastic nature of Diffusion-based completion introduces inconsistencies between synthesized and observed textures, this phenomenon becomes more prominent when the task involves multi-style faces and the dataset scale is limited.
- Naïve Poisson fusion occasionally produces tonal mismatches or unnatural blending boundaries under diverse input conditions.

To address these issues, we hypothesize that a comprehensive facial texture space could produce a refined texture T_{opt} that retains proximity to T_{init} while achieving superior style consistency and face feature alignment.

Currently, there are two main backbones to construct the facial texture space: Diffusion-based and GAN-based. While diffusion-based image generation models demonstrate remarkable generative quality, as discussed in DiffMorpher[36], the latent space of diffusion models is constructed from noise maps lacking semantic information. This leads to critical challenges when performing interpolation or optimization in this space. Consequently, such operations frequently result in abrupt content flickering or convergence to local minima, thereby preventing further refinement towards accurate results.

Instead, we construct the latent space for multi-style facial texture with a StyleGAN2[14] backbone. In particular, we train a StyleGAN2 network G_{gan} on our TexHub. We optimize within the W latent space of StyleGAN2 to identify an initial latent code W_{init} that minimizes the L1 distance between StyleGAN-generated texture T_{opt} and T_{init}. Then, the optimized texture T_{opt} is mapped to the mesh M and projected onto the image plane of the input I, generating an image I_M. We compute a compound loss combining pixel-level L1 norm and VGG-based perceptual loss between I_M and I,

$$\begin{aligned}\mathcal{L}_{\text{total}} =& \lambda_{\text{L1}} \cdot \|I_M - I\|_2^2 + \lambda_{\text{perc}} \cdot \|\phi(I_M) - \phi(I)\|_2^2 \\ &+ \lambda_{\text{reg}} \|W - W_{init}\|\end{aligned} \tag{1}$$

where ϕ denotes VGG-19 feature extraction. This loss drives iterative updates to the latent code, ultimately yielding the optimal W and its corresponding facial texture T_{opt}.

3.2.3 Texture Quality Enhancement

While T_{opt} exhibits strong semantic alignment with input image I, texture blurring persists due to the limited training data for StyleGAN2. To enhance high-frequency details, we further refine T_{opt} using a Diffusion-based repainting strategy. Leveraging our established facial texture generation workflow, we apply SDEdit-based[25] image-to-image translation by injecting noise (strength=0.3) into the VAE latent of T_{opt}. During Diffusion sampling, we enforce UV layout consistency through integrated LoRA adapters and canny-edge-guided ControlNet constraints, mirroring our texture data synthesis workflow. This produces the final high-fidelity texture map.

4. Experiments

4.1. Implementation Details

For the training of the diffusion-based inpainter G_{diff}, we train a StableDiffusion 2.1 network from scratch with 20,000 UV-texture images in *TexHub*. We train the network with batch size 16 and learning rate $5 * 10^{-5}$ (Adam), running for 90 epoches. For the training of the stylegan-based texture generator G_{gan}, we train a StyleGAN2 network from scratch with 20,000 UV-texture images in *TexHub*. We train the network with batch size 32 and learning rate 0.0025 for both generator and discriminator, running for 10,000 iterations on four NVIDIA GeForce RTX 3090 Ti GPUs.

The optimization on G_{gan} is divided into two stages. For the optimization of the initialization latent, we first run on the Z latent space for 100 steps, then 500 steps on the W latent space. For the optimization of texture correction, we run on the W latent space for 100 steps. All of the optimizations are performed with Adam of 0.001 learning rate.

4.2. Comparison

To demonstrate the effectiveness of our proposed framework, we compare our AvatarTex with the methods directly

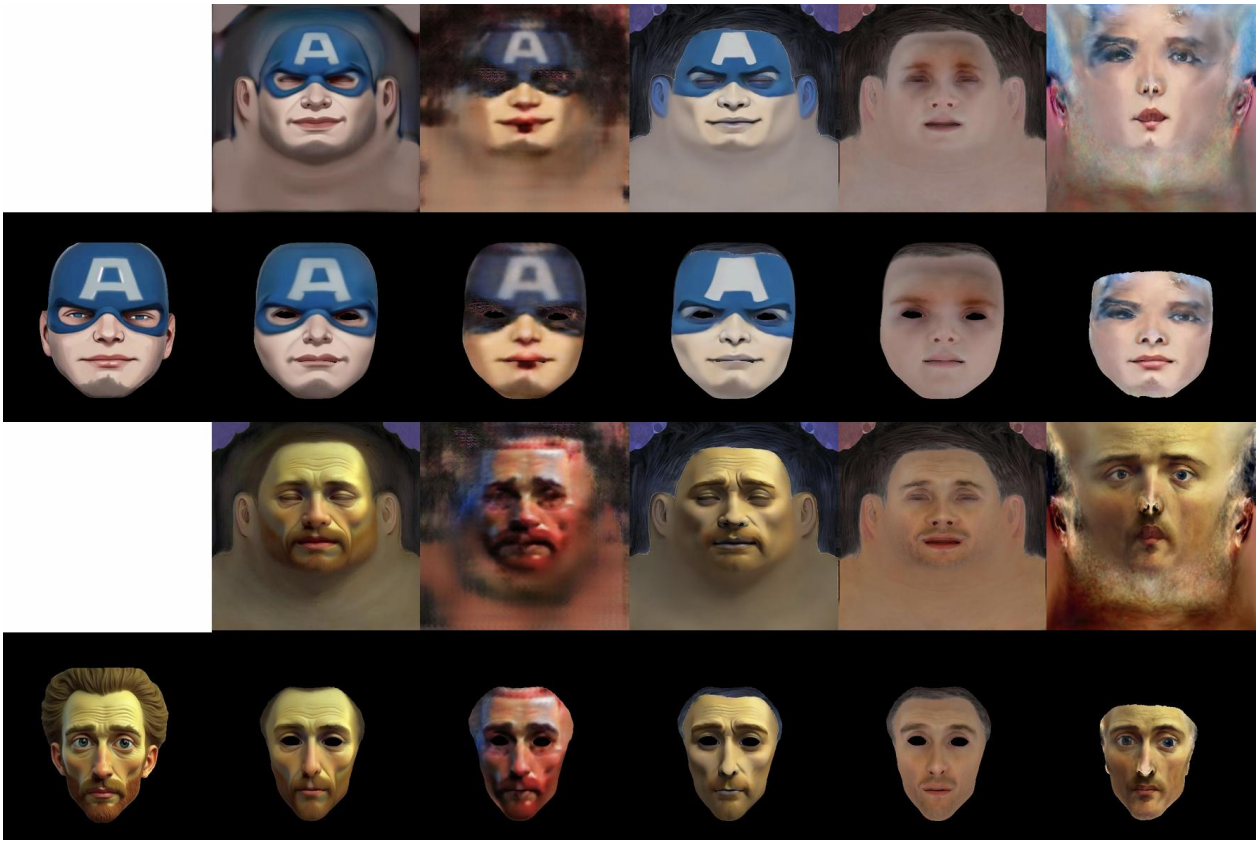

Figure 5. The visual results of our comparisons. The results are (a) ours (b) pixel2pixelHD inpainting[32] (c) Stable Diffusion 2.1[30] inpainting (d) FFHQ-UV[1] (e)UV-IDM[21] respectively. More examples can be found in the gallery of the supplementary material.

performing texture inpainting on T_{proj} , thereby highlighting the necessity of texture space optimization for acquiring high-quality texture images. For texture image completion, we construct two baseline approaches: one based on Pixel2PixelHD[32] and another utilizing Stable Diffusion 2.1[30]. We train both inpainting networks on our *TexHub* dataset. Additionally, we conducted comparative analysis with current state-of-the-art real-world facial texture reconstruction methods FFHQ-UV[1] and UV-IDM[21]. Fig. 5 presents the results of texture reconstruction results with these methods. As shown in Fig. 5(b), the trained Pixel2PixelHD struggles to achieve reasonable completion for texture images in the test set. Fig. 5(c) reveals that while the trained Stable Diffusion 2.1 can generate textures with a coherent overall appearance, significantly discrepancies exist in facial features and texture characteristics compared to the target image. Fig. 5(d,e) show models trained on real facial data demonstrate limited generalizability when directly applied to multi-style texture images. Fig. 5(a) shows our AvatarTex effectively reconstructs the texture features from target images.

4.3. Ablation Study

We designed comprehensive ablation studies to validate the effectiveness of individual components within our proposed framework. Our texture reconstruction methodology consists of three stages: texture initialization, texture content consistency optimization, and texture detail quality enhancement. We established six experimental configurations as shown in Tab. 1. The results are shown in Fig. 6. Fig. 6(a) shows that without a proper initialization, the optimization on G_{gan} tends to generate blurry results. In comparison, Fig. 6(d) shows the optimization on G_{diff} could generate a texture image with sharp details but fails to capture the local face feature, as shown in the second case. When optimizing from T_{init}, both of G_{gan} and G_{diff} obtain better UV texture, as shown in Fig. 6(b,e). The results from G_{gan} capture more accurate facial features than the ones from G_{diff}. For example, the eyebrow tattoo of the first case in column (b) matches the input better than the one in column (e). But still the low-level details in Fig. 6(b) are blurry, due to the capacity of trained StyleGAN2. The results (our full method) in Fig. 6(c) achieve high-quality

Figure 6. The visual results of our ablation studies. The setting of each column is shown in Tab. 1.

topology-aligned texture.

4.4. Quantitative Evaluation

We conduct quantitative experiments to compare our method with SOTA approaches and perform an ablation study on our pipeline. For evaluation, we collect a test dataset of 100 in-the-wild face images sourced from public websites. We assess texture reconstruction performance using four metrics: PSNR, SSIM, LPIPS, and FID. The experimental results are presented in Tab. 2 and Tab. 3.

Setting	T_{init}	Latent Space	Enhancement
a		StyleGAN	
b	✓	StyleGAN	
c	✓	StyleGAN	✓
d		SD 2.1	
e	✓	SD 2.1	
f	✓	SD 2.1	✓

Table 1. The configuration for our ablation studies.

	PSNR↑	SSIM↑	LPIPS↓	FID↓	ID↑	User Study↑
Pix2PixHD	21.38	0.89	0.41	60.75	0.51	1.88
FFHQ-UV	22.53	0.89	0.36	49.63	0.54	1.56
UV-IDM	24.25	0.92	0.25	39.89	0.59	2.75
SD2.1	25.18	0.93	0.23	31.08	0.60	2.94
Ours	**30.65**	**0.96**	**0.16**	**21.46**	**0.81**	**4.62**

Table 2. The quantitative results on comparison.

Setting			Metric					
T_{init}	Prior	Enhance	PSNR↑	SSIM↑	LPIPS↓	FID↓	ID↑	User Study↑
	StyleGAN		28.57	0.94	0.22	30.37	0.61	2.91
	SD 2.1		23.49	0.92	0.29	32.13	0.60	1.79
✓	StyleGAN		29.41	0.95	0.20	24.73	0.80	3.92
✓	SD 2.1		23.88	0.93	0.28	31.46	0.64	2.54
✓	StyleGAN	✓	**30.65**	**0.96**	**0.16**	**21.46**	**0.81**	**5.29**
✓	SD 2.1	✓	23.95	0.93	0.28	31.47	0.62	2.54

Table 3. The quantitative results on ablation.

4.5. User Study

We further conduct a user study evaluating visual quality, shown in sixth column of the tables. For this study, we randomly select 20 test samples and ask participants to rank the results of each method based on perceived visual quality. The highest possible score is 5 in Tab. 2, and 6 in Tab. 3.

5. Conclusion and Limitation

We present AvatarTex, a facial texture reconstruction framework combining diffusion models with GAN-based structural regularization for high-fidelity, topology-consistent textures across stylized and photorealistic faces. We introduce TexHub, a 20K multi-style UV texture dataset that improves generalization. AvatarTex outperforms prior methods in style fidelity, detail preservation, and geometric coherence. TexHub will be released to support future research. AvatarTex does not explicitly disentangle shading and albedo, which may limit relightability.

References

[1] Haoran Bai, Di Kang, Haoxian Zhang, Jinshan Pan, and Linchao Bao. Ffhq-uv: Normalized facial uv-texture dataset for 3d face reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 362–371, 2023. 2, 3, 7

[2] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 4

[3] 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 *IEEE Computer Vision and Pattern Recognition Workshops*, 2019. 3

[4] Abdallah Dib, Junghyun Ahn, Cedric Thebault, Philippe-Henri Gosselin, and Louis Chevallier. S2f2: Self-supervised high fidelity face reconstruction from monocular image. *arXiv preprint arXiv:2203.07732*, 2022. 3

[5] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4): 1–13, 2021. 3

[6] Yuming Gu, Hongyi Xu, You Xie, Guoxian Song, Yichun Shi, Di Chang, Jing Yang, and Linjie Luo. Diffportrait3d: Controllable diffusion for zero-shot portrait view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10456–10465, 2024. 3

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[8] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022. 2

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 4

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 2

[11] Wonjong Jang, Yucheol Jung, Hyomin Kim, Gwangjin Ju, Chaewon Son, Jooeun Son, and Seungyong Lee. Toonify3d: Stylegan-based 3d stylized face generator. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[12] Yucheol Jung, Wonjong Jang, Soongjin Kim, Jiaolong Yang, Xin Tong, and Seungyong Lee. Deep deformable 3d caricatures with learned shape control. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 3

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020. 2, 6

[15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 3

[16] Tobias Kirschstein, Simon Giebenhain, Jiapeng Tang, Markos Georgopoulos, and Matthias Nießner. Gghead: Fast and generalizable 3d gaussian heads. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 3

[17] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2, 4

[18] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos P Zafeiriou. Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[19] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. Fitme: Deep photorealistic 3d morphable model avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8629–8640, 2023. 3

[20] 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. 2023. 3

[21] Hong Li, Yutang Feng, Song Xue, Xuhui Liu, Bohan Zeng, Shanglin Li, Boyu Liu, Jianzhuang Liu, Shumin Han, and Baochang Zhang. Uv-idm: identity-conditioned latent diffusion model for face uv-texture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10585–10595, 2024. 2, 3, 4, 7

[22] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2

[23] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2

[24] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9970–9980, 2024. 2

[25] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 5, 6

[26] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 577–582. 2023. 5

[27] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[28] Yuda Qiu, Xiaojie Xu, Lingteng Qiu, Yan Pan, Yushuang Wu, Weikai Chen, and Xiaoguang Han. 3dcaricshop: A dataset and a baseline method for single-view 3d caricature face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10245, 2021. 3

[29] Xingyu Ren, Jiankang Deng, Yuhao Cheng, Jia Guo, Chao Ma, Yichao Yan, Wenhan Zhu, and Xiaokang Yang. Monocular identity-conditioned facial reflectance reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 885–895, 2024. 3

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2, 4, 7

[31] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 2

[32] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 4, 7

[33] Zidu Wang, Xiangyu Zhu, Tianshuo Zhang, Baiqin Wang, and Zhen Lei. 3d face reconstruction with the geometric guidance of facial part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1672–1682, 2024. 3

[34] Haoyu Wu, Ziqiao Peng, Xukun Zhou, Yunfei Cheng, Jun He, Hongyan Liu, and Zhaoxin Fan. Vgg-tex: A vivid geometry-guided facial texture estimation model for high fidelity monocular 3d face reconstruction. *arXiv preprint arXiv:2409.09740*, 2024. 3

[35] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2, 3, 4

[36] Kaiwen Zhang, Yifan Zhou, Xudong Xu, Xingang Pan, and Bo Dai. Diffmorpher: Unleashing the capability of diffusion models for image morphing. *arXiv preprint arXiv:2312.07409*, 2023. 6

[37] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 4

[38] Mingyuan Zhou, Rakib Hyder, Ziwei Xuan, and Guojun Qi. Ultravatar: A realistic animatable 3d avatar diffusion model with authenticity guided textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1238–1248, 2024. 2, 4