PaintHuman: Towards High-Fidelity Text-to-3D Human Texturing via Denoised Score Distillation

Jianhui Yu[1], Hao Zhu[2], Liming Jiang[3], Chen Change Loy[3], Weidong Cai[1], Wayne Wu[2]
[1]University of Sydney
[2]Shanghai AI Laboratory
[3]S-Lab, Nanyang Technological University
jianhui.yu@sydney.edu.au, haozhu96@gmail.com, liming002@ntu.edu.sg,
ccloy@ntu.edu.sg, tom.cai@sydney.edu.au, wuwenyan0503@gmail.com

Abstract

Recent advances in zero-shot text-to-3D human generation, which employ the human model prior (*e.g.*, SMPL) or Score Distillation Sampling (SDS) with pre-trained text-to-image diffusion models, have been groundbreaking. However, SDS may provide inaccurate gradient directions under the weak diffusion guidance, as it tends to produce over-smoothed results and generate body textures that are inconsistent with the detailed mesh geometry. Therefore, directly leveraging existing strategies for high-fidelity text-to-3D human texturing is challenging. In this work, we propose a model called *PaintHuman* to addresses the challenges from two perspectives. We first propose a novel score function, *Denoised Score Distillation (DSD)*, which directly modifies the SDS by introducing negative gradient components to iteratively correct the gradient direction and generate high-quality textures. In addition, we use the depth map as a geometric guide to ensure that the texture is semantically aligned to human mesh surfaces. To guarantee the quality of rendered results, we employ geometry-aware networks to predict surface materials and render realistic human textures. Extensive experiments, benchmarked against state-of-the-art (SoTA) methods, validate the efficacy of our approach. **Project page:** https://painthuman.github.io/

Introduction

Significant progress has been made in text-to-3D content generation. Some methods are proposed for general objects (Poole et al. 2023; Lin et al. 2023), and some are specifically for 3D human avatars (Cao et al. 2023; Hong et al. 2022; Jiang et al. 2023). 3D human avatars are increasingly important in various applications, including games, films, and metaverse. In this work, we focus on texturing a predefined human mesh with text prompts.

The success of recent methods rely on CLIP model (Hong et al. 2022) or text-to-image generation, which leverages the diffusion model (Ho, Jain, and Abbeel 2020; Rombach et al. 2022), and Score Distillation Sampling (SDS) (Poole et al. 2023) combined with differentiable 3D representations (Mildenhall et al. 2020; Barron et al. 2021). However, directly leveraging existing strategies for detailed human avatar texturing in a zero-shot manner is challenging for two reasons. First, we find that SDS is a general-purpose optimization, which guides the loss gradient in a direction due to its weak supervision and unable to well handle unclear signal from the diffusion model. This issue results in generated human textures of low quality, including over-smoothed body parts and blurry garment details. Second, textures guided by text-to-image models are usually not semantically unaligned to either input texts or human mesh surfaces, resulting in missing textures or unaligned texture mapping for the geometry.

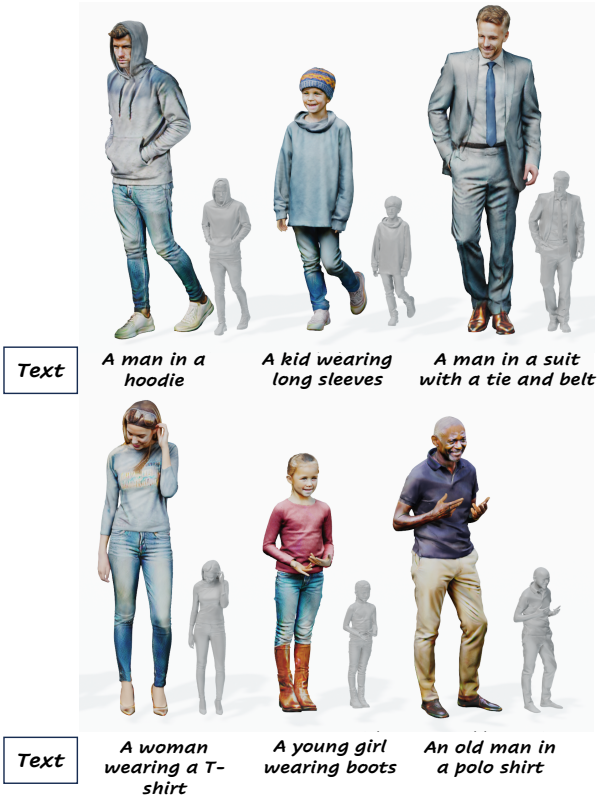


Figure 1: Generated results of PaintHuman. Given textureless human meshes and textual descriptions as input, our model can generate high-quality and detailed textures that aligned to input geometry and texts.

Recent work (Hong et al. 2022; Jiang et al. 2023) for human avatar texturing entangles shape and texture generation, which leverage human-specific priors (Loper et al. 2015) for human body texturing. To ensure the generated textures aligned to the given geometry, TEXTure (Richardson et al. 2023) and Text2Tex (Chen et al. 2023a) utilize a depth-aware diffusion model (Rombach et al. 2022) to directly in-paint and update textures from different viewpoints, which could cause inconsistency when the input mesh has complex geometry. Other methods such as Latent-Paint (Metzer et al. 2023) or Fantasia3D (Chen et al. 2023b) apply SDS to update the loss gradient for consistent texture generation. However, SDS fails to semantically align textures to input texts, *i.e*., the synthesized textures are non-detailed and over-smoothed.

Therefore, we propose *PaintHuman* to address a primary issue associated with SDS. Our main idea is to denoise the unclear gradient direction provided by the SDS loss. We handle this from two aspects. Firstly, we propose *Denoising Score Distillation* (DSD), which introduces a negative gradient component to directly modify the SDS, which could iteratively correct the gradient direction for detailed and high-quality texture generation. Then, to enable geometry-aware texture generation, we utilize geometric guidance which provides rich details of the mesh surface to guide the DSD precisely, and use spatially-aware texture shading models (Karis and Games 2013) to guarantee the quality of rendered visual results.

Specifically, DSD utilizes an additional negative pair of image and text. The key idea is that by using a negative image, *i.e*., an image with noise rendered from the last training iteration, we could reinforce the learning of the complex surface geometry to produce clear boundaries between different garments. In addition, with the help of negative text prompts, the synthesized textures could be more semantically aligned to the input text. Overall, the negative pair contributes a negative part to SDS, which controls the gradient direction by a weighted subtraction of the two input pairs, producing an effective gradient to address over-smoothed texture generation. To further ensure textures semantically aligned to the complex avatar surface, we first use the depth map as guidance during the diffusion process for texturing, which provides fine-grained surface details. In addition, we follow (Munkberg et al. 2022) to apply the Spatially-Varying Bidirectional Reflectance Distribution Function (SV-BRDF) (Karis and Games 2013) and coordinate-based networks (Müller et al. 2022) for geometry-aware material prediction. With the help of differentiable rendering (Hasselgren et al. 2021), we could update the rendered human avatar and synthesized textures in an end-to-end fashion.

The contributions of our work are summarized as follows:

- We introduce Denoising Score Distillation (DSD), a diffusion-based denoising score using negative image-text pairs for high-fidelity texture generation aligned to textual descriptions.
- We employ semantically aligned 2D depth signals and spatially-aware rendering functions for geometry-aware texture generation and realistic avatar rendering.
- Through comprehensive experiments, we prove the efficacy of our method over existing texture generation techniques.

Related Work

Diffusion Models. With the development of denoising score-matching generative models (Sohl-Dickstein et al. 2015), diffusion models present great success in a variety of domains such as image editing, text-to-image synthesis, text-to-video synthesis, and text-to-3D synthesis. In the field of text-to-image synthesis, diffusion models have demonstrated impressive performance, especially the Stable Diffusion model (Rombach et al. 2022), which is trained on a large number of paired text-image data samples with CLIP (Radford et al. 2021) to encode text prompts and VQ-VAE (Van Den Oord, Vinyals et al. 2017) to encode images into latent space. In our work, we use a pre-trained Stable Diffusion model to incorporate intrinsic image prior to guide the training of our texture generation network.

3D Shape and Texture Generation. There has been a recent surge of interest in the field of generating 3D shapes and textures. One line of methods, such as Text2Mesh (Michel et al. 2022), Tango (Lei et al. 2022), and CLIP-Mesh (Mohammad Khalid et al. 2022), utilize CLIP-space similarities as an optimization objective to create novel 3D shapes and textures. Gao et al. (2022) trains a model to generate shape and texture via a DMTet (Shen et al. 2021) mesh extractor and 2D adversarial losses. A recent approach called DreamFusion (Poole et al. 2023) introduces the use of pre-trained diffusion models to generate 3D NeRF (Mildenhall et al. 2020) models based on a given text prompt. The key component in DreamFusion is the score distillation sampling (SDS), which uses a pre-trained 2D diffusion model as a critique to minimize the distribution of the predicted and ground-truth Gaussian noise, thus the 3D scene can be optimized for desired shape and texture generation.

In the context of texture generation, Latent-NeRF (Metzer et al. 2023) demonstrated how to employ SDS loss in the latent space of the diffusion model to generate textures for 3D meshes and then decoded to RGB for the final colorization output. Besides, both TEXTure (Richardson et al. 2023) and Text2Tex (Chen et al. 2023a) proposed a non-optimization method with progressive updates from multiple viewpoints to in-paint the texture over the 3D mesh models.

Human-specific shape and texture generation methods also follow the same ideas that use CLIP similarity between the generated human image and the textural descriptions (Hong et al. 2022) or directly leverage SDS for iterative shape and texture generation (Kolotouros et al. 2023; Zeng et al. 2023; Jiang et al. 2023). Besides, they also employ human body model prior, *i.e*., SMPL (Loper et al. 2015), for effective human avatar generation. However, most generated human textures are over-smooth and of low quality, which we argue is caused by the unstable guidance provided by SDS.

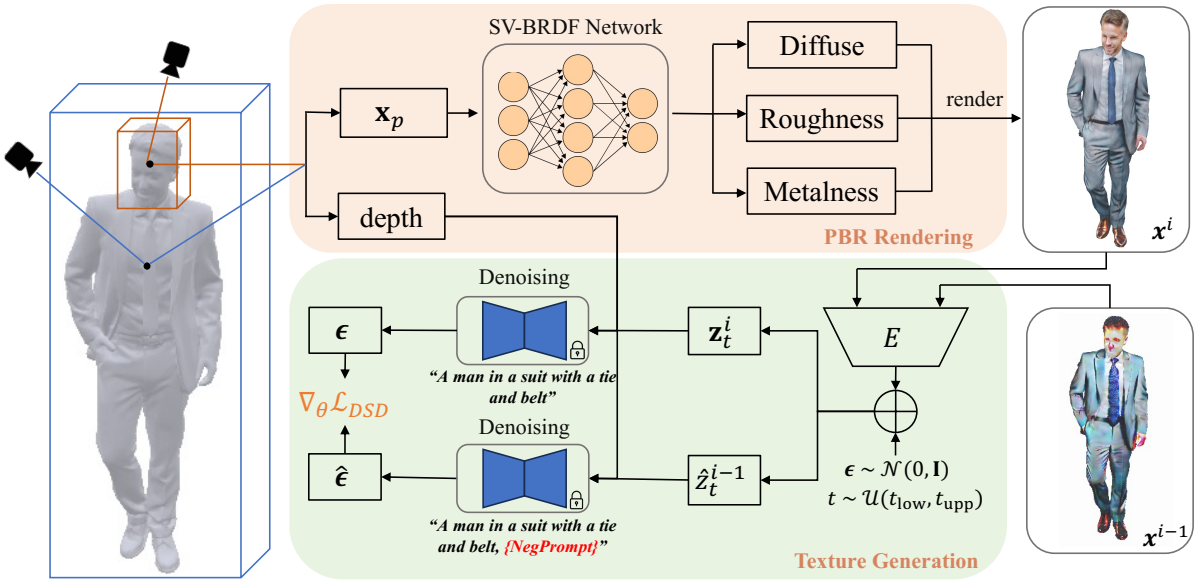


Figure 2: Overview of our proposed model. Our goal is to texture the human mesh given an input text and a mesh model. To achieve this, we propose Denoised Score Distillation with a negative pair of image and text prompts to guide the gradient direction for detail texture generation that is semantically aligned to the input text. We introduce depth signals to the diffusion process for complex garment texturing, and a learnable network to estimate SV-BRDFs for albedo and material parameter learning. Finally, the camera position is adjusted for refined details of the face region.

Method

In this section, we start with an overview of SDS. We then introduce Denoised Score Distillation (DSD), which uses an extra negative pair of image-text to guide gradient direction, thereby generating detailed textures that align with the input text. Finally, we employ depth signals in the diffusion process for complex surface texturing and employ a geometry-aware rendering function for photorealistic human texture generation. The overall pipeline is shown in Figure 2.

SDS Overview

Given an input image $\mathbf{x}$ with a latent code $\mathbf{z}$, a conditioning text embedding y, a denoising U-Net ϵ_ϕ with model parameters ϕ, a uniformly sampled timestep $t \sim \mathcal{U}(0, \mathbf{I})$, and a Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, the diffusion loss is:

$$\mathcal{L}_{\text{Diff}}(\mathbf{z}, y, t) = w(t)\|\epsilon_\phi(\mathbf{z}_t, y, t) - \epsilon\|_2^2, \tag{1}$$

where $w(t)$ is a weighting function depending on t, and $\mathbf{z}_t$ refers to the noisy version of $\mathbf{z}$ via an iterative forward diffusion process given by $\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z} + \sqrt{1-\alpha_t}\epsilon$, with α_t being the noise scheduler. For high-quality generation, classifier-free guidance (CFG) (Ho and Salimans 2022) is used, which jointly learns text-conditioned and unconditioned models via a scale parameter ω. During inference, the two models are used to denoise the image as follows:

$$\hat{\epsilon}_\phi(\mathbf{z}_t, y, t) = (1+\omega)\epsilon_\phi(\mathbf{z}_t, y, t) - \omega\epsilon_\phi(\mathbf{z}_t, t). \tag{2}$$

Given a differentiable rendering function g_θ, the gradient of diffusion loss with respect to model parameters θ is:

$$\nabla_\theta \mathcal{L}_{\text{SDS}} = w(t)\left(\hat{\epsilon}_\phi(\mathbf{z}_t, y, t) - \epsilon\right)\frac{\partial \mathbf{z}_t}{\partial \theta}, \tag{3}$$

where we have omitted the U-Net Jacobian term as shown in (Poole et al. 2023). The purpose of SDS is to generate samples via optimization from a text-guided diffusion model. However, we argue that SDS only presents poor guidance on input text prompt and the generated 2D image, hence, in the following, we propose a new loss design to increase the generation quality.

Denoised Score Distillation

Given a textureless human avatar, our task is to generate surface textures conditioned on input texts. Due to SDS and neural representation of 3D avatar (Mildenhall et al. 2020), zero-shot human texture generation is made possible. We observe that using SDS only for human texturing can cause over-smoothed body parts and cannot be fully semantically aligned to the input text.

We address the issue brought by SDS by proposing a new method, Denoised Score Distillation (DSD), for detailed human avatar texturing of high quality. Specifically, when presented with input text embedding y and the corresponding image $\mathbf{x}$ with the latent code $\mathbf{z}$, our objective is to refine the gradient $\nabla_\theta \mathcal{L}_{\text{SDS}}$ in Eq. 3 to a direction, so that the rendered avatar contains a detailed texture mapping that is semantically aligned to the input text. Mathematically, our DSD score function is formulated as:

$$\mathcal{L}_{\text{DSD}} = w(t)\left(\|\epsilon_\phi(\mathbf{z}_t^i, y, t) - \epsilon\|_2^2 - \lambda\|\epsilon_\phi(\hat{\mathbf{z}}_t^{i-1}, \hat{y}, t) - \epsilon\|_2^2\right), \tag{4}$$

where we introduce a *negative* pair of image with latent code $\hat{\mathbf{z}}$ and text with embedding $\hat{y}$. λ is a weighting parameter. Both $\mathbf{z}_t^i$ and $\hat{\mathbf{z}}_t^{i-1}$ have a superscript i indicating the training iteration and share the same timestep t and noise ϵ, allow-

ing us to use the same U-Net for noise prediction. Then the gradient of $\mathcal{L}_{\text{DSD}}$ over the model parameter θ is:

$$\begin{aligned}\nabla_\theta \mathcal{L}_{\text{DSD}} &= w(t)\big(\hat{\epsilon}_\phi\left(\mathbf{z}_t, y, t\right) - \epsilon - \lambda(\hat{\epsilon}_\phi\left(\hat{\mathbf{z}}_t, \hat{y}, t\right) - \epsilon)\big)\frac{\partial \mathbf{z}_t}{\partial \theta} \\ &= w(t)\big(\hat{\epsilon}_\phi\left(\mathbf{z}_t, y, t\right) - \lambda\hat{\epsilon}_\phi\left(\hat{\mathbf{z}}_t, \hat{y}, t\right) - (1-\lambda)\epsilon\big)\frac{\partial \mathbf{z}_t}{\partial \theta},\end{aligned} \tag{5}$$

where we have omitted the U-Net Jacobian matrix following Poole et al. (2023).

As shown in Figure 2, we employ the negative image $\hat{\mathbf{x}}^{i-1}$ derived from the previous training iteration, where we consider $\hat{\mathbf{x}}^{i-1}$ a negative version of $\mathbf{x}^i$ as it contains more noise signals. The inclusion of the negative image within the computation process of $\nabla_\theta \mathcal{L}_{\text{DSD}}$ yields two significant advantages. Firstly, $\hat{\mathbf{z}}_t^{i-1}$ can reinforce the memory of the rendered human image during training, so that the final output can still be semantically aligned to the input text. Secondly, the incorporation of the negative image improves the model's capacity to learn complex geometries, thus facilitating the generation of clear boundaries between varying garment types. For negative prompts, we use common prompts such as *disfigured*, *ugly*, etc. However, we would adapt existing prompts based on a test run, infusing refined negative prompts based on the observed output. For instance, if artifacts emerge within rendered hand regions, we append "*bad hands*" to the prompt set. In contrast to the indirect application of negative prompts in Stable Diffusion, we inject the negative prompt embedding directly into $\nabla_\theta \mathcal{L}_{\text{DSD}}$. This strategy effectively minimizes the artifact in rendered human images, thereby enhancing the quality of the generated output.

Through the integration of both negative image and prompts, we successfully manipulate the existing SDS gradient in Eq. 3 to guide the model convergence towards a mode that yields highly detailed and qualitative textures, which also remain semantically aligned to the input text. Further analyses and insights into this approach are provided in our ablation study.

Geometry-aware Texture Generation

Geometry Guidance in DSD. To accurately texture complex garment details, we compute and leverage the corresponding depth map as a fine-grained guidance. Therefore, we employ a pre-trained depth-to-image diffusion model (Rombach et al. 2022) rather than the general version, so that the generated avatar could follow the same depth values of the given surface mesh. As shown in Figure 5 (b), although the rendered human image presents textures that are not semantically aligned to the input text as the belt region is not clearly textured, utilizing the depth-aware diffusion model ensures the generated texture reserve more geometric details and semantically aligned to the given geometry.

Shading Model for Rendering. Following the idea of physically based rendering (PBR), which models and renders real-world light conditions and material properties, we estimate surface materials by leveraging SV-BRDFs for human image rendering:

$$R(\mathbf{x}_p, \mathbf{l}) = \int_H L_i(\mathbf{l})(f_d + f_s)\,(\mathbf{l}\cdot\mathbf{n})\,\mathrm{d}\mathbf{l}, \tag{6}$$

where $L_i(\mathbf{l})$ is the incident radiance, and $H = \{\mathbf{l} : \mathbf{l}\cdot\mathbf{n} \geq 0\}$ denotes a hemisphere with the incident light and surface normal $\mathbf{n}$. f_s and f_d are diffuse and specular SV-BRDFs, respectively.

In particular, we follow (Karis and Games 2013) to employ a simple diffuse model at a low cost. The diffuse SV-BRDF is mathematically expressed as follows:

$$f_d(\mathbf{x}_p) = \frac{\mathbf{k}_d}{\pi}, \tag{7}$$

where $\mathbf{k}_d$ is the diffuse term which can be learned based on 3D vertex positions. For specular SV-BRDF estimation, we use a microfacet specular shading model as in (Karis and Games 2013) to characterize the physical properties of the mesh surface:

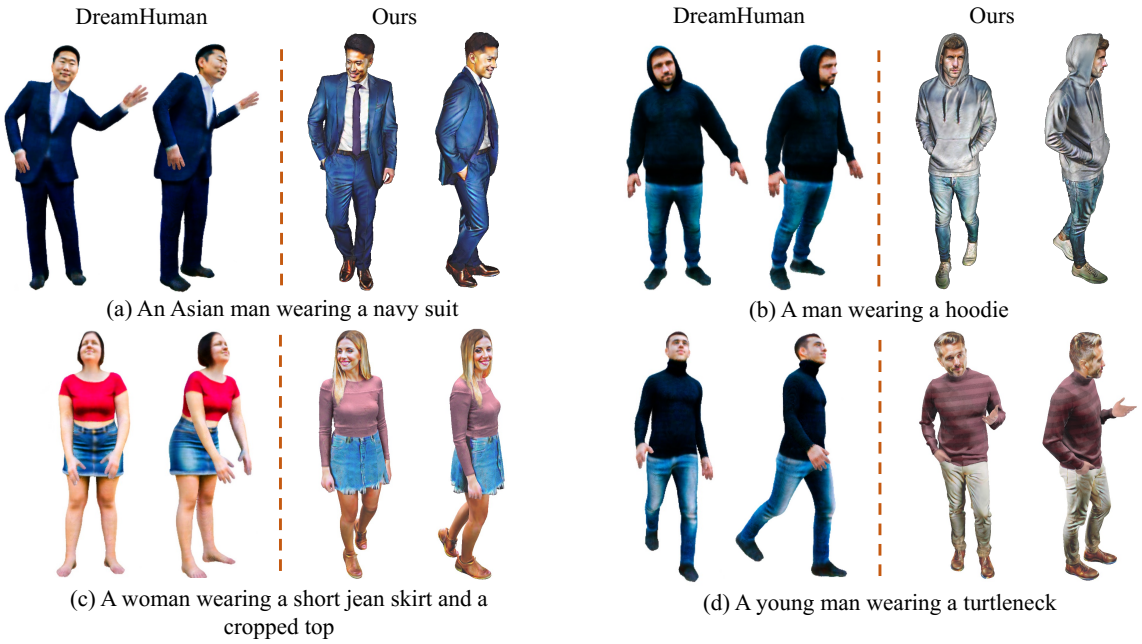
$$f_s(\mathbf{l}, \mathbf{v}) = \frac{DFG}{4(\mathbf{n}\cdot\mathbf{l})(\mathbf{n}\cdot\mathbf{v})}, \tag{8}$$

where $\mathbf{v}$ is the view direction. D, F and G represent the normal distribution function, the Fresnel term and geometric attenuation, respectively. We also choose the Disney BRDF Basecolor-Metallic parametrization (Burley and Studios 2012) for a physically accurate rendering. Specifically, the specular reflectance term $\mathbf{k}_s = m\cdot\mathbf{k}_d + (1-m)\cdot 0.04$, where the diffuse term $\mathbf{k}_d$, the roughness term r and the metallic term m can be estimated via our proposed SV-BRDF network give surface points $\mathbf{x}_p$: $\sigma(\gamma(\mathbf{x}_p)) = [\mathbf{k}_d, r, m]$. γ is a coordinate-based network (Müller et al. 2022) and σ is a parameterized SV-BRDF estimation model. We also utilize a differentiable split-sum approximation for Eq. 6 to maintain the differentiability in the rendering process. Moreover, we follow (Zhang et al. 2021) to regularize the material learning, which results in a smooth albedo map.

Semantic Zoom. Human perception is particularly sensitive to distortions and artifacts in facial features. However, texturing human avatars in a full-body context often results in degraded facial details. To address this issue, we enhance the human prior during the optimization process by semantically augmenting the prompt (Hong et al. 2022). For example, we prepend "the face of" to the beginning of the prompt to pay more attention to this region. Simultaneously, every four iterations, we shift the look-at point of the camera to the face center and semantically zoom into the facial region, which refines facial features and improves the overall perception of the rendered avatar.

Experiments

Baseline Methods. We compare our model to recent SoTA baseline models, including Latent-Paint (Metzer et al. 2023), TEXTure (Richardson et al. 2023), and Fantasia3D (Chen et al. 2023b) with the appearance modeling part only. We modify Fantasia3D to ensure the vertex positions remain fixed. We also compare our model to a recent method for realistic human avatar generation, DreamHuman (Kolotouros et al. 2023), to further validate the effectiveness of



(a) An Asian man wearing a navy suit

(b) A man wearing a hoodie

(c) A woman wearing a short jean skirt and a cropped top

(d) A young man wearing a turtleneck

Figure 3: Qualitative comparisons with DreamHuman (Kolotouros et al. 2023). As DreamHuman is not publicly available, we pick similar mesh models from Renderpeople (Renderpeople 2021) and download the results from the published paper.

our design. Although its human mesh model is not publicly available, we use the same text prompts as DreamHuman to evaluate the texture quality with similar mesh models.

Qualitative Analysis. As shown in Figure 4, we compare our results against baseline models. Latent-Paint cannot capture the object semantics, which results in failed or blurry textured avatars. TEXTure generates relatively better results than Latent-Paint but suffers from inconsistent textures. Fantasia3D performs well given certain input texts as in Figure 4 (a) and (c), but it outputs unrealistic samples with noisy textures in most cases due to the use of SDS. In contrast, our model produces textured avatars with high-quality and detailed textures, which are aligned to input texts and consistent with the geometry. We compare our model with DreamHuman in Figure 3. We observe that using the same text input, our model generates textured avatars with more high-frequency details, such as the cloth wrinkles, which is different from DreamHuman where the textures are over-smoothed. Moreover, in both experiments, our model can consistently generate high-quality human faces.

Quantitative Analysis. To investigate the alignment between the rendered human avatars and the input texts, we use the CLIP score (Radford et al. 2021). As shown in Table 1, we compare our method with the baseline models and report the mean CLIP score. Specifically, we generate 6 frontal images from all textured avatars, each separated by a 30-degree interval. We use 20 different meshes with 4 prompts for each mesh, with a total of 80 prompts. We observe that our model outperforms all baseline models, where our result is higher than Latent-Paint by the largest margin of around 19.99%.

Method	Mean CLIP Score	Δ (%)
Latent-Paint	24.11	19.99%
TEXTure	25.34	14.17%
Fantasia3D	27.10	6.75%
PaintHuman (Ours)	**28.93**	-
DreamHuman	25.79	12.25%
PaintHuman (Ours)	**28.95**	-

Table 1: Quantitative comparisons between baseline models and ours. Δ denotes the percentage by which our model outperforms the indicated method.

Such improvements demonstrate that our proposed DSD is capable of generating more realistic textures on complex human meshes, and is better aligned to the input texts.

User Study. We conduct user study to analyze the quality of the generated textures and the fidelity to the input text. Specifically, 4 meshes are selected with 4 text prompts generated for each mesh, resulting in 16 visual results for each baseline method. We ask users to rate the overall quality, including texture quality and alignment between text and rendered results. More details are shown in the supplementary material. Collected results are reported in Table 2 including mean scores and standard deviation values, indicating that our method outperforms the baselines.

Ablation Study. To validate the effectiveness of our design, we use two prompts as examples: "*a man in a suit with a belt and tie*" and "*a young man wearing a turtleneck*" for the ablation study. Results are shown in Figures 5 and 6.

Figure 4: Qualitative comparisons on RenderPeople (Renderpeople 2021) for textured human avatars. Compared with Latent-Paint (Metzer et al. 2023), TEXTure (Richardson et al. 2023), and Fantasia3D (Chen et al. 2023b), our generations contain the best texture quality with high-frequency details and consistent with input textual descriptions.

(a) SDS (b) SDS + Depth (c) SDS + Depth + NegPrompt (d) DSD + SH (f) Full: DSD + Depth + BRDF

Figure 5: Visualization of ablation study. We provide results of textured human avatars based on different settings that are added gradually from vanilla SDS baseline.

Method	Score	Δ (%)
Latent-Paint	1.21±0.70	148.76%
TEXTure	1.28±0.60	135.16%
Fantasia3D	1.76±0.70	71.02%
DreamHuman	2.83±0.82	6.36%
PaintHuman (Ours)	**3.01±0.95**	-

Table 2: User study results of baseline models and ours. Δ denotes the percentage by which our model outperforms the indicated method.

Firstly, the efficacy of our DSD is verified through several comparisons. As shown in Figure 5(a), we note that employing SDS for human texturing often results in over-smoothed body parts and fails to fully align with the input text semantically, where the belt region is neglected. The addition of depth map guidance in Figure 5(b) also struggles to address this issue. Furthermore, by adding negative prompts, Figure 5(c) shows that the rendered image contains more high-frequency details but is not aligned with the input text, and some parts are devoid of texturing.

We further examine the effectiveness of the BRDF shading model. As shown in Figure 5(d), we render the result with the Spherical Harmonic model (SH) (Boss et al. 2021), resulting in less realistic textures with noticeably noisy color distributions at the borders between different garments. However, using BRDF can give us smooth and clear textures. In contrast, as shown in Figure 5(e), an image rendered using our DSD effectively mitigates the over-smoothing issue and results in a detailed human avatar.

Finally, as shown in Figure 6, our application of semantic zoom on the face region enhances the overall texture quality. Notably, the method enables the presence of intricate facial features resulting in a more realistic representation.

Figure 6: Importance of semantic zoom. The left image shows the generated avatar with semantic zoom, while the right image employs no semantic zoom.

Conclusion

In this work, we introduce PaintHuman, a zero-shot text-to-human texture generation model. We present a novel score function, Denoised Score Distillation (DSD), which refines the gradient direction to generate high-quality, detailed human textures aligned to the input text. We also leverage geometry signals in DSD for accurate texturing of complex garment details. To maintain semantic alignment between the mesh and the synthesized texture, we employ a differentiable network to parameterize SV-BRDFs for surface material prediction, which is complemented by physically based rendering for realistic avatar renderings, with facial details refined through semantic zooming. Our extensive experiments reveal significant improvements in texture generation, validating the effectiveness of our module designs.

References

Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5855–5864.

Boss, M.; Braun, R.; Jampani, V.; Barron, J. T.; Liu, C.; and Lensch, H. 2021. Nerd: Neural reflectance decomposition from image collections. In *ICCV*.

Burley, B.; and Studios, W. D. A. 2012. Physically-based shading at disney. *Siggraph*.

Cao, Y.; Cao, Y.-P.; Han, K.; Shan, Y.; and Wong, K.-Y. K. 2023. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. *arXiv preprint arXiv:2304.00916*.

Chen, D. Z.; Siddiqui, Y.; Lee, H.-Y.; Tulyakov, S.; and Nießner, M. 2023a. Text2Tex: Text-driven texture synthesis via diffusion models. *ICCV*.

Chen, R.; Chen, Y.; Jiao, N.; and Jia, K. 2023b. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *ICCV*.

Gao, J.; Shen, T.; Wang, Z.; Chen, W.; Yin, K.; Li, D.; Litany, O.; Gojcic, Z.; and Fidler, S. 2022. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances In Neural Information Processing Systems*, 35: 31841–31854.

Hasselgren, J.; Munkberg, J.; Lehtinen, J.; Aittala, M.; and Laine, S. 2021. Appearance-Driven Automatic 3D Model Simplification. In *EGSR*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hong, F.; Zhang, M.; Pan, L.; Cai, Z.; Yang, L.; and Liu, Z. 2022. AvatarCLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars. *ACM ToG*.

Jiang, R.; Wang, C.; Zhang, J.; Chai, M.; He, M.; Chen, D.; and Liao, J. 2023. AvatarCraft: Transforming Text into Neural Human Avatars with Parameterized Shape and Pose Control. *arXiv preprint arXiv:2303.17606*.

Karis, B.; and Games, E. 2013. Real shading in unreal engine 4. *Proc. Physically Based Shading Theory Practice*.

Kolotouros, N.; Alldieck, T.; Zanfir, A.; Bazavan, E. G.; Fieraru, M.; and Sminchisescu, C. 2023. DreamHuman: Animatable 3D Avatars from Text. *arXiv preprint arXiv:2306.09329*.

Lei, J.; Zhang, Y.; Jia, K.; et al. 2022. Tango: Text-driven photorealistic and robust 3d stylization via lighting decomposition. *Advances in Neural Information Processing Systems*, 30923–30936.

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2023. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 300–309.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A skinned multi-person linear model. *ACM ToG*.

Metzer, G.; Richardson, E.; Patashnik, O.; Giryes, R.; and Cohen-Or, D. 2023. Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures. In *CVPR*.

Michel, O.; Bar-On, R.; Liu, R.; Benaim, S.; and Hanocka, R. 2022. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13492–13502.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

Mohammad Khalid, N.; Xie, T.; Belilovsky, E.; and Popa, T. 2022. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 conference papers*, 1–8.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM ToG*.

Munkberg, J.; Hasselgren, J.; Shen, T.; Gao, J.; Chen, W.; Evans, A.; Müller, T.; and Fidler, S. 2022. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8280–8290.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *ICLR*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Renderpeople. 2021. https://renderpeople.com/. Accessed: 2023-07-21.

Richardson, E.; Metzer, G.; Alaluf, Y.; Giryes, R.; and Cohen-Or, D. 2023. TEXTure: Text-Guided Texturing of 3D Shapes. In *SIGGRAPH*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Shen, T.; Gao, J.; Yin, K.; Liu, M.-Y.; and Fidler, S. 2021. Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis. In *NeurIPS*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*.

Zeng, Y.; Lu, Y.; Ji, X.; Yao, Y.; Zhu, H.; and Cao, X. 2023. AvatarBooth: High-Quality and Customizable 3D Human Avatar Generation. *arXiv preprint arXiv:2306.09864*.

Zhang, X.; Srinivasan, P. P.; Deng, B.; Debevec, P.; Freeman, W. T.; and Barron, J. T. 2021. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM ToG*.