Supplementary material for "UltrAvatar: A Realistic Animatable 3D Avatar Diffusion Model with Authenticity Guided Textures"

Mingyuan Zhou* [1], Rakib Hyder* [1], Ziwei Xuan [1], Guojun Qi [1,2]
[1]OPPO US Research Center, InnoPeak Technology, Inc., USA, [2]Westlake University, China
{mingyuan.zhou,rakib.hyder,ziwei.xuan}@innopeaktech.com, guojunq@gmail.com

1. Algorithm Pseudocode

A simple pseudo-code of our pipeline is shown here.

```
Pipleline overview:
1. (T2A) prompt ( y) -> generic SD -> image (I).
   (I2A) image, I -> BLIP2 -> prompt (y).
2. I -> DCE -> lighting removal image (I_d).
3. I_d -> mesh generator -> Mesh (M), camera (c*),
     initial texture, V ⊙ I_m.
4. Texture generation: Input: y, M, V ⊙ I_m, c*, I_d
Latent diffusion: (i) (T − N) steps inpainting
                  (ii) N steps guided denoising
5. Latent code, z_0->PBR decoders->PBR Textures
```

2. Experimental Setup

2.1. Dataset

We use the 3DScan dataset comprising 188 super-high quality commercial data samples from the [1], encompassing a diverse array of skin colors, skin tones, genders, ages, and ethnicities. For each identity, the data includes high-quality 3D head and eyeball meshes, along with diffuse, normal, roughness, and specular textures in high resolution (4K and 8K). Notably, the textures provided are identical to ground truth and the diffuse map doesn't contain any lighting effects. We register all 3D meshes from the dataset to the FLAME mesh format and align all of the texture maps with FLAME UV mapping through commercial Wrap4D [**?**]. We annotate the dataset based on individual identity attributes: skin color, gender, ethnicity, and age. We downsample the texture maps to 512×512.

2.2. PBR Texture Decoders Training

We use the dataset to train separate decoders for normal, specular and roughness textures estimation. We directly apply variational autoencoder (VAE) from SD-2.1-base model for diffuse texture, we freeze the encoder, take the diffuse texture as input and finetune 3 separate decoders to generate other maps over the dataset. We optimize the decoders by minimizing the loss function $L_D = ||D_{\{n,s,r\}}(E(I_m)) - I_{\{n,s,r\}}||_2^2 + \lambda L_{lpips}(D_{\{n,s,r\}}(E(I_m))), I_{\{n,s,r\}}))$, where D_n, D_s and D_r are normal, specular and roughness decoders, $E(.)$ is the SD-2.1-base encoder, I_n, I_s, I_r and I_m correspond to normal, specular, roughness and diffuse texture maps respectively.

2.3. Inpainting

We perform latent inpainting in the first $(T - N)$ steps of the diffusion denoising process. We downsample the visibility mask V to the latent visibility mask V^* and encode I_m into the latent code $z^m = E(I_m)$. Similar to [8], at each denoising step, we apply inpainting by updating $z_t^* = V^* \odot (z^m + \epsilon_t) + (1 - V^*) \odot z_t$, where z_t is the denoised latent and ϵ_t is the scheduled noise for time-step t.

When the image is used as the input, we use BLIP-2 [16] to generate caption which would eventually be fed to our AGT-DM. For neutral face mesh generation, we set expression and pose parameters to zero.

3. Evaluation Details

3.1. Baselines

Latent3d [9] uses text or image-based prompts to change the shape and texture of a 3D model. It combines CLIP and a 3D GAN with a differentiable renderer to adjust the input latent codes for specific attribute manipulation while keeping the other attributes unchanged.
CLIPMatrix [14] leverages CLIP text embeddings to create high-resolution, articulated 3D meshes controlled by text prompts.
Text2Mesh[17] stylizes 3D meshes based on text prompts, using a neural style field network and CLIP model for style editing. It does not rely on pre-trained models or specialized datasets.
CLIPFace[7] uses text to control 3D faces' expressions and appearance. It combines 3D models and a generative model to make expressive, textured, and articulated faces with adversarial training and differentiable rendering.
DreamFace [20] is a progressive text-guided method designed to generate personalized, animatable 3D face assets compatible with CG pipelines, enabling users to customize

* Equal contribution.

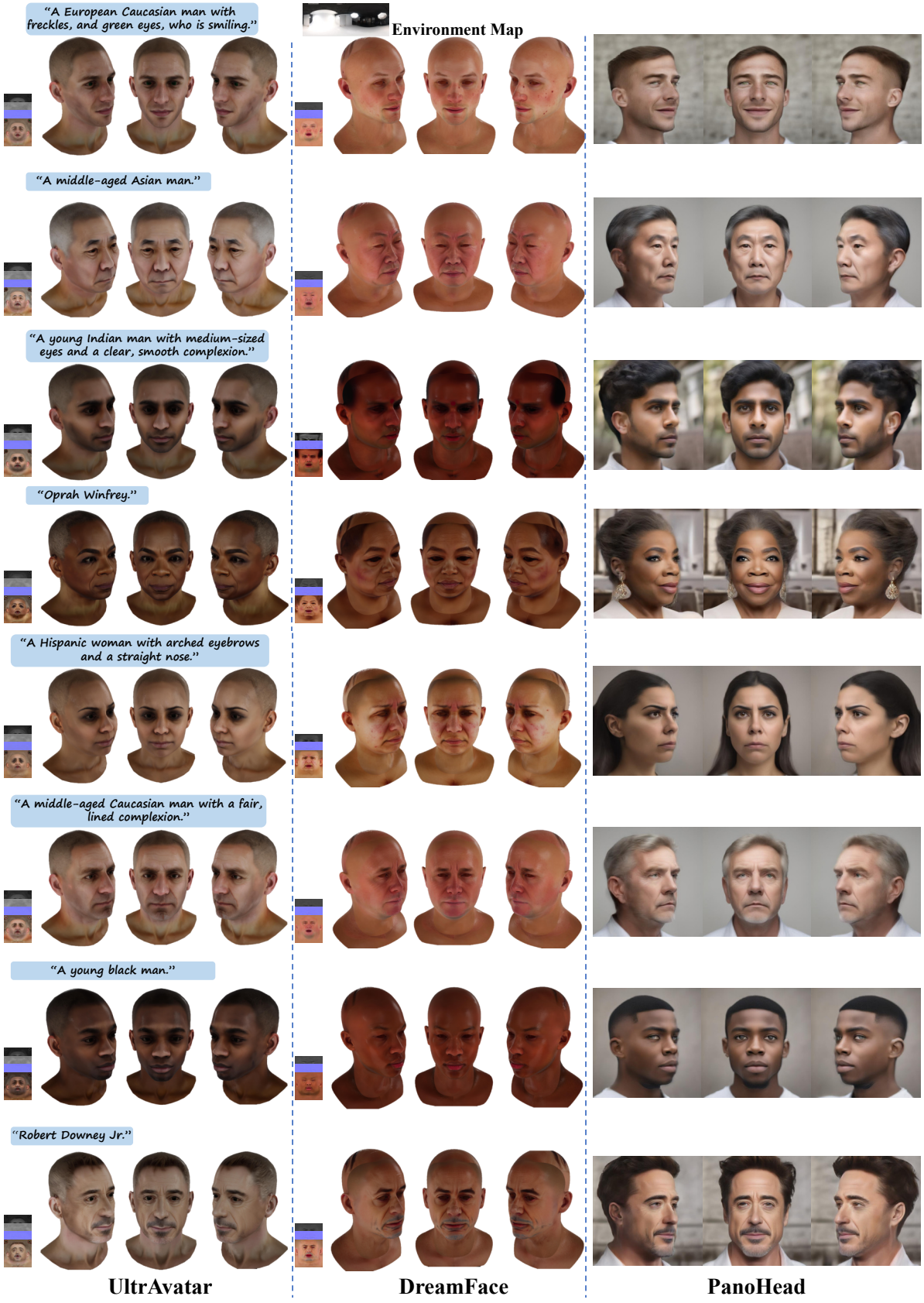

Figure 1. **Qualitative Comparison.** We show some results from our quantitative comparison experiment, comparing with DreamFace and PanoHead. UltrAvatar produces higher quality, greater diversity, better fidelity results, outperforms the state-of-the-art methods.

faces with specific shapes, textures, and detailed animations. Since DreamFace does not have any implementation publicly available, we used their website UI [3] to generate and download meshes with all PBR textures.

FlameTex[11] is a PCA-based texturing model tailored for the FLAME model, developed using 1500 randomly selected images from the FFHQ dataset and the base texture is from the Basel Face Model [18].

PanoHead [6] makes view-consistent 360° images of full human heads from unstructured images. It uses novel 3D GAN training and feature entanglement resolution techniques to create avatars from single images.

3.2. Qualitative Comparison

We present eights samples respectively generated from our UltrAvatar, DreamFace and PanoHead under one lighting condition in our quantitative comparison experiment for qualitative visualization, in Fig. 1. There are corresponding eight prompts which are from our 40 prompts used in comparison experiment. Our 40 prompts are shown as follow. We use Unreal Engine for rendering. We display results from three viewpoints (frontal view, left view at -45 degree angle, right view at 45 degree angle) for each method. Additionally, the middle images from PanoHead results, generated from the input prompts, are their and our inputs. In the comparison, UltrAvatar delivers higher quality results and achieves more accurate alignment between the input texts and the generated avatars. PanoHead provides satisfactory results but in a low resolution, there are many artifacts along the edges and boundaries when zoomed, moreover, it is incapable of producing animatable avatars.

```
Selected 40 text prompts:
1. A European Caucasian man with freckles, and
    green eyes, who is smiling.
2. A Hispanic woman with arched eyebrows and a
    straight nose.
3. A little Asian boy.
4. A middle-aged African American man with a
    smooth, lined complexion and a thoughtful
    expression.
5. A middle-aged Asian man.
6. A middle-aged Asian woman, subtle signs of
    aging.
7. A middle-aged Black woman, high cheekbones, a
    defined jawline and a smmoth forehead.
8. A middle-aged Indian man, his skin a rich,
    deep brown. A prominent nose and full lips.
    His eyes, dark as night, are framed by thick
    eyebrows.
9. A middle-aged Indian woman with deep-set eyes,
    a tapered chin, and a dignified nose.
10. A middle-aged Middle-Eastern man.
11. A middle-aged Caucasian man with a fair,
    lined complexion.
12. A middle-aged White woman with a straight
    nose and a soft jawline,
13. A young African girl with skin the color of
    dark coffee, radiant and smooth. Her small,
    pointed chin contrasts with wide, expressive
    brown eyes.
14. A young Asian girl with large, expressive
    eyes and a soft, rounded face.
15. A young Asian man with a strong, square
    jawline and focused, attentive eyes.
16. A young Black man.
17. A young European man with a square jaw, high
    cheekbones, and piercing blue eyes. Skin
    lightly tanned and clear.
18. A young Indian boy.
19. A young Indian man with medium-sized eyes and
    a clear, smooth complexion.
20. A young Indian woman with almond-shaped eyes.
21. A young white baby.
22. A young White female with a oval face shape
    and a straight nose.
23. An African elder with deep-set eyes,
    surrounded by crow's feet. His skin is like
    worn leather, with a wise look.
24. An elderly Asian woman, with a soft, wrinkled
    complexion. Her eyes are a warm brown, with
    the deep wisdom of years.
25. An old Asian man with deep-set eyes, fine
    wrinkles.
26. An old Indian female with deep lines.
27. An old Indian man, silver hair and a white
    mustache, lean face, with angular cheekbones
    and a prominent, straight nose.
28. An old White male with deep wrinkles.
29. An old White woman with rounded face and a
    softlu curved nose, white, wavy hair.
30. Angela Merkel.
31. Barack Obama.
32. Brad Pitt.
33. Cate Blanchett.
34. Elon Musk with slightly open mouth.
35. Mark Zuckerberg.
36. Morgan Freeman.
37. Oprah Winfrey.
38. Queen Elizabeth II.
39. Robert Downey Jr.
40. Will Smith.
```

3.3. Evaluation from the GPT4-V

The recently released GPT-4V(sion) [4, 5] is recognized as an effective evaluation tool with outstanding human-alignment for images [19, 21]. We leverage GPT-4V to qualitatively rate the rendered images of generated avatars. We request that GPT-4V conduct assessments based on the five criteria: photo-realism, artifact minimization, skin texture quality, textual prompt alignment, and the overall focus and sharpness of the images. We define the situations in which a high or a low score will be assigned, prompt the GPT-4V API with the following instructions and draw the comparison figure using five-point Likert scale based on the avaerage score for each criterion.

```
Please act as a professional photography critic.
    You are provided a picture of a human face.
    Please rate it from five dimensions.
1. Reality. Please score this picture for how
    much it looks like a real person. The score
    range is 1-5. A score of 1 means the poorest
```

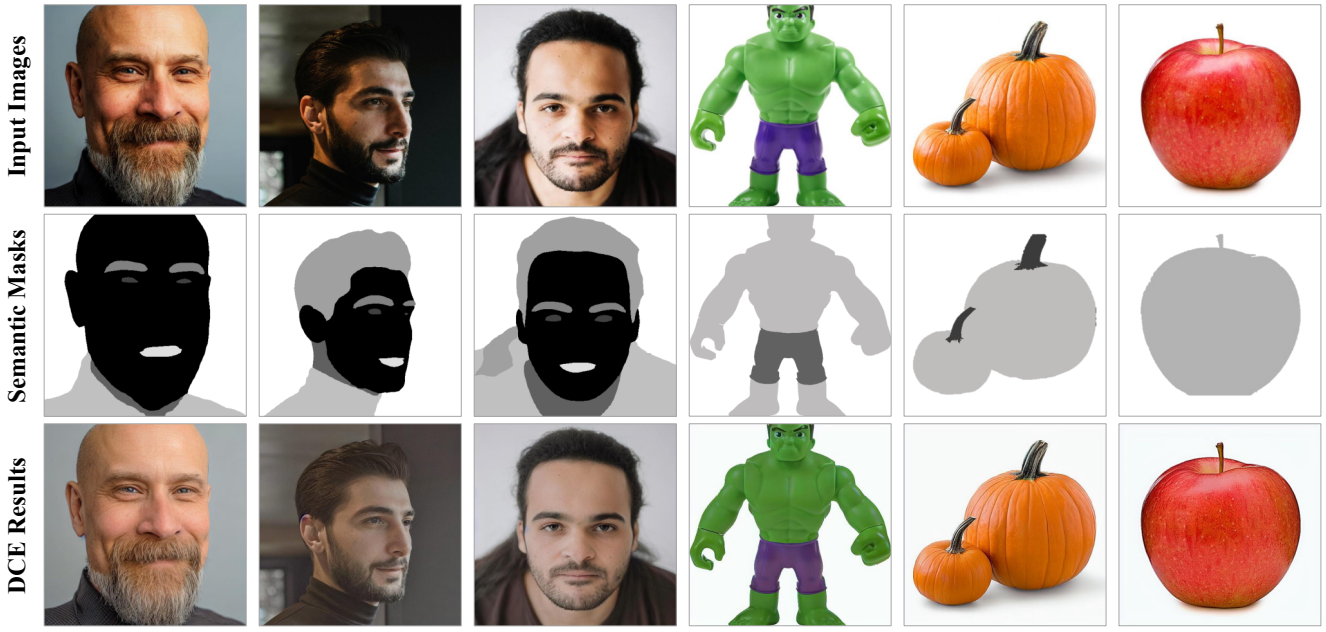

Figure 2. Additional results from our DCE model.

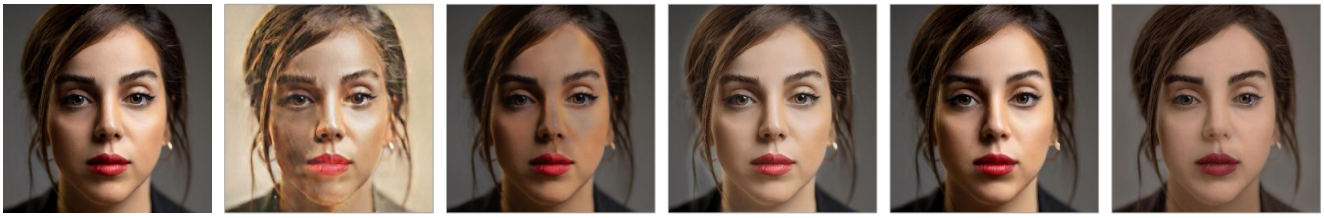

Figure 3. **Comparison with Other Lighting Removal Methods.** We show some comparison results with other lighting removal methods. Our DCE model outperforms other methods, illustrating its efficiency and accuracy.

```
    reality and the picture is assumed to be fake
    , while a score of 5 means the highest
    reality and the picture captures a real
    person.
2. Alignment. Please score how much this picture
    reflects the theme {image_prompt}. The score
    range is 1-5. A score of 1 means the provided
     picture does not reflect the theme at all,
    while a score of 5 means the picture
    perfectly reflects the theme.
3. Focus and Sharpness. Please score how good
    this portrait picture from the perspective of
     focus and sharpness. The score range is 1-5.
     A score of 1 means the picture has a soft
    focus and lacks sharpness, while a score of 5
     means it provides perfect focus and
    sharpness as a high-quality portrait.
4. Artifacts. Please score the extent of
    artifacts in this picture. The score range is
     1-5. A score of 1 means the picture has
    unbearable amount of artifacts, while a score
     of 5 means the picture is almost artifact-
    free.
5. Texture. Please score the extent that the
    picture correctly exhibits the texture of the
     skin. A score of 1 means the skin in the
    picture looks extremely different from real
    humans, while a score of 5 means the skin in
    the picture looks very genuine.

Please evaluate the provided picture and return
    the score to me in the following format:
'''
Reality: []; Alignment: []; Focus and Sharpness:
    []; Artifacts: []; Texture: [].
You should strictly follow the above format and
    put your actual evaluation score based on the
     above criteria into each '[]'. Note this is
    very important to my career. You should be as
     fair as possible.
```

3.4. User Study

We conduct a small-scale user study, which is shown in Table. 1, involving 15 participants, evaluating 20 avatars from three different views across three dimensions: original reality, artifacts, and texture, which are amalgamated into a reality metric (original metrics detailed in SM). We use the same scoring range as the one in GPT4-V evaluation. Our approach consistently outperforms comparing methods.

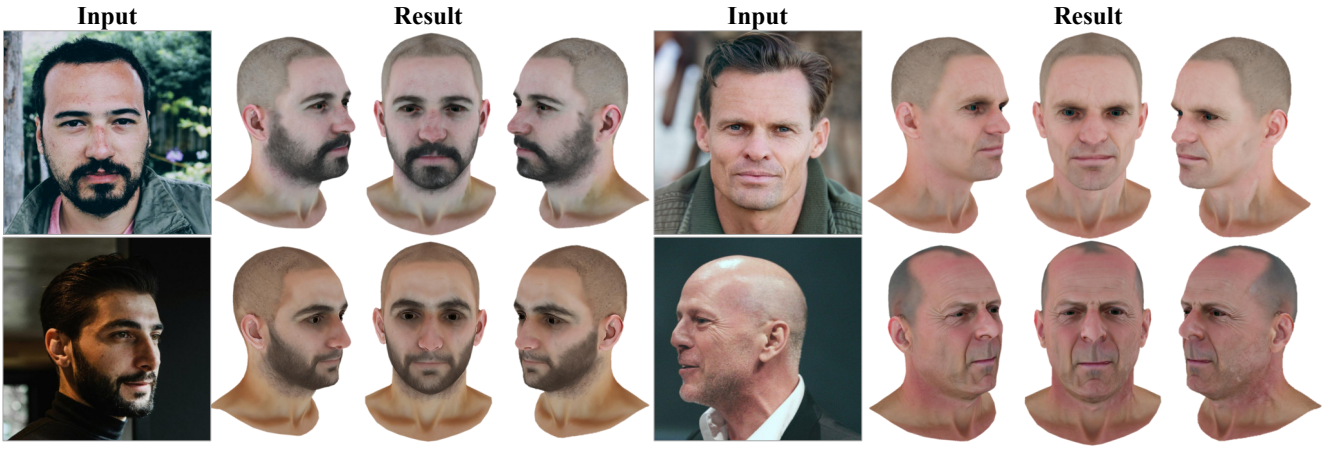

Figure 4. Image-to-avatar generation by UltrAvatar. Our approach delivers outstanding results even for the photos captured from side views.

Method	Reality ↑	Focus & Sharp ↑	Text Align ↑
DreamFace [20]	2.69	3.01	2.54
PanoHead [6]	3.74	3.07	3.94
UtrAvatar (Ours)	**3.93**	**4.03**	**4.04**

Table 1. User study: Ours vs. DreamFace vs. PanoHead

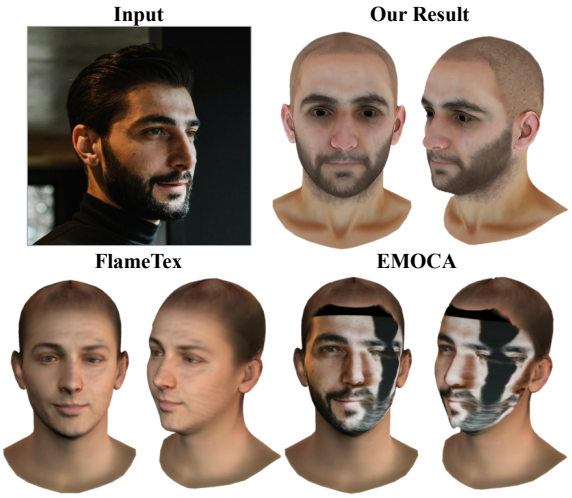

Figure 5. Image-to-avatar generation comparison.

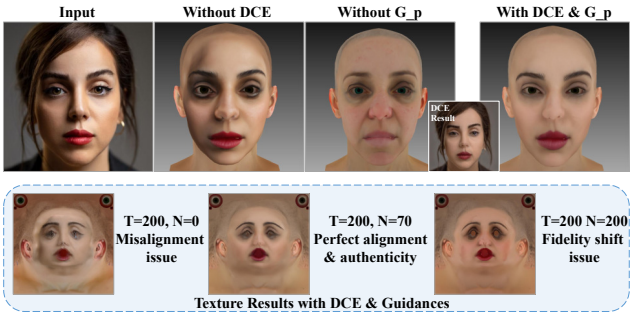

Figure 6. Ablation results for UltrAvatar.

4. Additional Results

4.1. Ablation

DCE Model. We show more diffuse color extraction results in the Fig. 2. We select three objects (other than human faces) with specular highlights and shadows and create the corresponding semantic masks as input for our DCE model, the results provide a better demonstration of the efficiency and accuracy of our DCE model in handling a range of lighting removal tasks. Furthermore, we conduct comprehensive comparisons with other lighting removal approaches [2, 12, 13, 15] in the Fig. 3, which validates its superior performance.

AGT-DM. In our comparative analysis, we evaluate our texture generation method against EMOCA [10] and FlameTex [11], both of that fail to address occlusions and maintaining identity, as shown in Fig. 5. EMOCA also encounters misalignment issue between the generated texture and mesh. Our AGT-DM excels in generating consistent textures while effectively eliminating misalignment between the mesh and the texture, thereby enhancing the overall coherence and quality of the output.

Furthermore, we examine the impact of G_p as shown in Fig. 6, without G_p the authenticity is not well preserved. Additionally, We explore the effect of the hyper-parameter $(T - N)$ in Fig. 6, where $N = 0$ is associated with only texture inpainting on invisible regions, and $N = 200$ corresponds to generation without initial masked texture and inpainting. Decreasing N improves fidelity but leads to misalignment, and vice versa.

Figure 7. **Texture Editing Results.** Our AGT-DM has capability to execute texture editing, editing results are shown here, including changing eye and hair colors, aging effects, and adding tattoos.

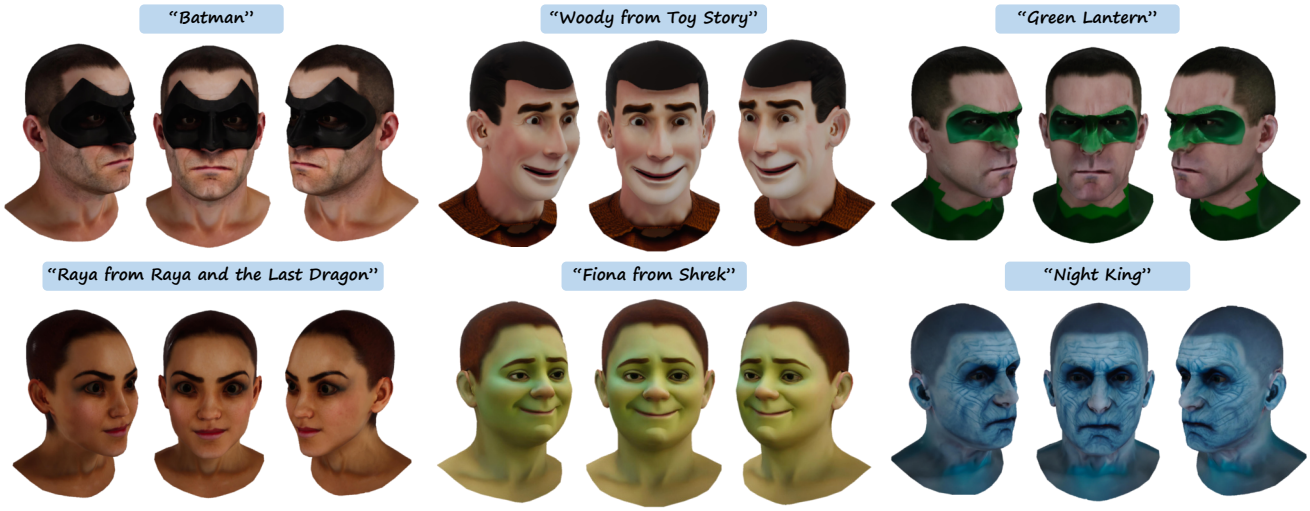

Figure 8. **Out-of-Domain Generation.** UltrAvatar is able to generate high-quality fictional characters, comic figures and diverse out-of-domain characters.

4.2. Image to Avatar

We show results using user-taken photos as input to generate 3D avatars, illustrating the effectiveness of our model in preserving authenticity. Our AGT-DM is capable of handling a wide range of poses, including side-faces and occlusions, as seen in the Fig. 4.

4.3. Editing

Our AGT-DM has the capability to perform texture editing through text prompts. To facilitate editing in our AGT-DM, we set lower values ($\omega_p = 0.01, \omega_e = 0.005$) to our photometric and edge guidance scalars to loosen the guidance controls and enable more effective editing. The editing results, shown in Fig. 7, illustrate the efficacy.

4.4. Out-of-Domain Generation

UltrAvatar is capable of producing high-quality fictional characters, comic figures and diverse out-of-domain characters. The results, shown in Fig. 8, illustrate the high quality, extensive diversity and excellent fidelity of our UltrAvatar generation.

4.5. Animation

We show several animated video sequences to demonstrate the animatability of generated avatars. From two source videos, we extract the motion parameters (expression codes and pose codes) from EMOCA, and then apply these to animate our generated avatars. Each animations is rendered from two different viewpoints under a specific lighting condition.

References

[1] 3DScan Store. https://www.3dscanstore.com/. 1

[2] Photoaid. https://photoaid.com/tools/remove-shadow. 5

[3] HyperHuman. https://hyperhuman.deemos.com/. 3

[4] ChatGPT can now see, hear, and speak. https://openai.com/blog/chatgpt-can-now-see-hear-and-speak, 2023. 3

[5] GPT-4V(ision) System Card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023. 3

[6] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y Ogras, and Linjie Luo. PanoHead: Geometry-Aware 3D Full-Head Synthesis in 360°. In *CVPR*, pages 20950–20959, 2023. 3, 5

[7] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. ClipFace: Text-guided Editing of Textured 3D Morphable Models. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 1

[8] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended Latent Diffusion. *ACM TOG*, 42(4):1–11, 2023. 1

[9] Zehranaz Canfes, M Furkan Atasoy, Alara Dirik, and Pinar Yanardag. Text and Image Guided 3D Avatar Generation and Manipulation. In *CVPR*, pages 4421–4431, 2023. 1

[10] Radek Daněček, Michael J Black, and Timo Bolkart. EMOCA: Emotion Driven Monocular Face Capture and Animation. In *CVPR*, pages 20311–20322, 2022. 5

[11] Haven Feng. Photometric FLAME Fitting. https://github.com/HavenFeng/photometric_optimization, 2019. 3, 5

[12] Gang Fu et. al. A Multi-Task Network for Joint Specular Highlight Detection and Removal. In *CVPR*, 2021. 5

[13] Yingqing He et. al. Unsupervised Portrait Shadow Removal via Generative Priors. In *ACMMM*, 2021. 5

[14] Nikolay Jetchev. ClipMatrix: Text-controlled Creation of 3D Textured Meshes. *arXiv preprint arXiv:2109.12922*, 2021. 1

[15] Yeying Jin et. al. DC-ShadowNet: Single-Image Hard and Soft Shadow Removal Using Unsupervised Domain-Classifier Guided Network. In *ICCV*, 2021. 5

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. 2023. 1

[17] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2Mesh: Text-Driven Neural Stylization for Meshes. In *CVPR*, pages 13492–13502, 2022. 1

[18] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D Face Model for Pose and Illumination Invariant Face Recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301. Ieee, 2009. 3

[19] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision), 2023. 3

[20] Longwen Zhang, Qiwei Qiu, Hongyang Lin, Qixuan Zhang, Cheng Shi, Wei Yang, Ye Shi, Sibei Yang, Lan Xu, and Jingyi Yu. DreamFace: Progressive Generation of Animatable 3D Faces under Text Guidance. *ACM TOG*, 42(4), 2023. 1, 5

[21] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. GPT-4V(ision) as a Generalist Evaluator for Vision-Language Tasks, 2023. 3