Supplementary Material for Paper: DiffusionGAN3D: Boosting Text-guided 3D Generation and Domain Adaptation by Combining 3D GANs and Diffusion Priors

Biwen Lei, Kai Yu, Mengyang Feng, Miaomiao Cui, Xuansong Xie
Alibaba Group
{biwen.lbw, jinmao.yk, mengyang.fmy, miaomiao.cmm}@alibaba-inc.com,
xingtong.xxs@taobao.com

In this document, we present full-text prompts used in the experiments in Sec. 1, implementation details in Sec. 2, more visualization results in Sec. 3, the application of DiffusionGAN3D on real images (3D-aware editing and stylization) in Sec. 4, discussions about the limitations of the proposed method and future work in Sec. 5.

1. Full-Text Prompts

In our paper, a shared additional positive prompt and the same negative prompt were applied in all the experiments including domain adaptation and text-to-avatar. The former was added after each primary prompt, and the latter served as a complete negative prompt for all experiments. Next, we provide the two shared prompts and all the primary prompts used in the experiments mentioned in the main paper and this document:

- additional positive prompt: ", sharp, 8K, skin detail, best quality, realistic lighting, good-looking, uniform light, extremely detailed"
- negative prompt: "blurry, inaccurate identity, disproportionate, wrong anatomy, blurry face, ugly, bad face lighting"

Main paper

- Figure 1.(Pixar): "Pixar style, a closeup of a person, cute, big eyes, Disney"
- Figure 1.(Greek statue): "a closeup of a Greek statue, head"
- Figure 1.(Joker): "joker, portrait" ("Zombie" in the 4th column of Figure 1. should be "Joker". It was a mistake.)
- Figure 1.(Pixar(cat)): "Pixar style, a closeup of a cat, Disney"
- Figure 1.(Fox, Pixar style): "Pixar style, a closeup of a fox, Disney"
- Figure 1.(Golden statue): "Golden statue, a close-up of a cat"
- Figure 1.(Hulk): "the Hulk, portrait"
- Figure 1.(Batman): "the Batman, portrait"
- Figure 1.(Obama): "Barack Obama, portrait"
- Figure 6.(Pixar): "Pixar style, a closeup of a person, Disney"
- Figure 6.(Lego): "Lego head"
- Figure 8.: "Pixar style, a cute girl with black hair"
- Figure 9.: "Pixar style, a closeup of a cat, Disney"
- Figure 10.: "a close-up of a woman with green hair"
- Figure 11.: "Link in Zelda, portrait"

This document

- Fig. 1.(Plaster statue): "a closeup of a plaster statue, head"
- Fig. 1.(Oil painting): "oil painting, a closeup of a person"
- Fig. 1.(Zombie): "zombie, head"
- Fig. 1.(Caricature): "caricature style, a closeup of a person"
- Fig. 1.(Cat, Pixar style): "Pixar style, a closeup of a cat, Disney"
- Fig. 1.(Fox, Pixar style): "Pixar style, a closeup of a fox, Disney"
- Fig. 2.(Purple hair): "a closeup of a person with purple hair"
- Fig. 2.(Blue eyes): "a closeup of a person with blue eyes"
- Fig. 2.(Red lipstick): "Red lipstick, a closeup of a person"
- Fig. 4.: "Catwoman"

2. Implementation Details

2.1. Architecture

Our model and baselines are implemented using Pytorch 1.13.1 and Python 3.8. We utilize EG3D-based [2] 3D GANs as our base generators, including PanoHead (head) [1], EG3D-FFHQ 512×512 (face), EG3D-AFHQ 512x512 (cat) and AG3D (body) [3]. We use the open-source Diffusers [12] library for the StableDiffusion [9] models. Training codes are built upon Stable-DreamFusion [11]. In the progressive texture refinement stage, we employ Py-

torch3D [7] as the differentiable rendering framework.

2.2. Training

The latent mapping network and the decoder are frozen during optimization. We only optimize the weights of the triplane generator for both domain adaptation and text-to-avatar generation tasks. We train the model using the Adam optimizer and a learning rate of 1×10^{-4}. All the models (both domain adaptation and text-to-avatar) are trained for 10,000 steps with a batch size of 1. Training takes about 50 minutes on an A100 GPU with a memory cost of 14G. Note that, in the domain adaptation task, we calculate the relative distance loss between the triplane results of the current batch and the previous batch. In SDS, the denoising timestep t is uniformly sampled from a range $T_{SDS} = (T_{min}, T_{max})$, where $T_{min} = 300$, $T_{max} = 800$ as default in our experiments. And the CFG [4] weight is set to 50 as default.

2.3. Progressive Texture Refinement

For 3D avatar generation task, we implement the texture refinement using uniformly selected $2k + 2$ ($2k$ indicates the views of the left and right sides, 2 denotes the front view and the back view) azimuths and j elevations. Specifically, we set $k = 1, j = 1$ for head generation, and $k = 2, j = 3$ for full-body generation. For the domain adaptation task, we only use 3 azimuths (-20, 0, 20 degrees) and a single elevation (0 degrees). The DDIM is adopted as the sampling method for the diffusion models and the number of denoising steps is set to 50. The denoising strengths of the image-to-image and inpainting are 0.6 and 0.4 respectively. The control weights for canny and depth conditions are both 1.

3. More Visualization Results

3.1. Domain adaptation

More comparison examples of 3D domain adaptation on EG3D-FFHQ(face) and EG3D-AFHQ(cat) are shown in Fig. 1. It can be seen that our method performs favorably against the others in terms of diversity, text-image correspondence, and texture quality. Please zoom in for a better view.

3.2. Local Editing

We compare the performance of our method and StyleGAN-Fusion [10] on some local editing scenarios. As shown in Fig. 2, StyleGAN-Fusion fails to preserve the details of the non-target region and leads to a global transformation. In this comparison, owing to the proposed diffusion-guided reconstruction loss, our method manages to precisely manipulate the target region while preserving the details of other regions and the overall identities.

3.3. Text-to-Avatar

We present more comparisons with other baselines on text-to-avatar tasks. As shown in Fig. 3, although based on the 3D generators trained on realistic human images, the proposed method is capable of generating avatars across large domain gaps, showing great generation capability and stability. By contrast, the text-to-3D methods suffer from convergence failure, over-saturation, and incorrect geometry. DreamHuman [5] also has the problem of over-smoothed texture.

Besides, we give an example of the results of the first stage of our method without performing texture refinement. As shown in Fig. 4, the generated geometry is smooth and the rendering results show decent quality and great view consistency.

4. Applications

Combined with the GAN inversion methods, DiffusionGAN3D is able to be applied to real face images to achieve text-guided 3D-aware stylization and local editing. We follow EG3D [2] and employ PTI [8] to project the real image into the latent space. Specifically, the PTI includes two steps: the latent code inversion and generator finetuning. The former step generates the inverted latent code w in $W+$ space and achieves rough reconstruction. The latter optimizes the generator and gets the finetuned generator G' for accurate recovering. In our framework, for the stylization task with large domain gaps, such as Pixar and Lego style, we directly input w into the stylized 3D GAN (such as the trained models in Fig. 1) to achieve 3D stylization. For the stylization tasks with a small domain gap or the local editing tasks, we need to re-implement the domain adaptation on G' to maintain the details and identity of the input image. Fig. 5 gives an example.

5. Limitations and Future Work

Limitations. We summarize two limitations of our method. On one hand, the performance of DiffusionGAN3D relies on the base 3D generator. We conducted extensive experiments on EG3D-FFHQ(face), PanoHead(head), and AG3D(body) for both domain adaptation and text-to-avatar tasks. The results show that the models trained on EG3D-FFHQ and PanoHead exhibit superior performance than the model trained on AG3D. A possible solution is to boost the 3D generator itself with a high-quality dataset generated by diffusion models. On the other hand, the proposed methods cannot well handle the local editing with deformation, such as the prompt "fat face". We assume that the diffusion model itself cannot deal with local deformation well, and therefore the SDS [6] cannot provide clear guidance for 3D generators.

Future Work. Beyond addressing the limitations discussed above, we will further extend our method to achieve accurate, high-fidelity and animatable avatar generation from images or text descriptions for future work, conquering some challenging problems (such as modeling complex geometry and appearance).

References

[1] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360deg. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20950–20959, 2023. 1

[2] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 1, 2, 4, 5

[3] Zijian Dong, Xu Chen, Jinlong Yang, Michael J Black, Otmar Hilliges, and Andreas Geiger. Ag3d: Learning to generate 3d avatars from 2d image collections. *arXiv preprint arXiv:2305.02312*, 2023. 1

[4] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[5] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *arXiv preprint arXiv:2306.09329*, 2023. 2

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

[7] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 2

[8] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021. 2

[9] 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*, pages 10684–10695, 2022. 1

[10] Kunpeng Song, Ligong Han, Bingchen Liu, Dimitris Metaxas, and Ahmed Elgammal. Diffusion guided domain adaptation of image generators. *arXiv preprint arXiv:2212.04473*, 2022. 2

[11] Jiaxiang Tang. Stable-dreamfusion: Text-to-3d with stable-diffusion, 2022. https://github.com/ashawkey/stable-dreamfusion. 1

[12] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 1

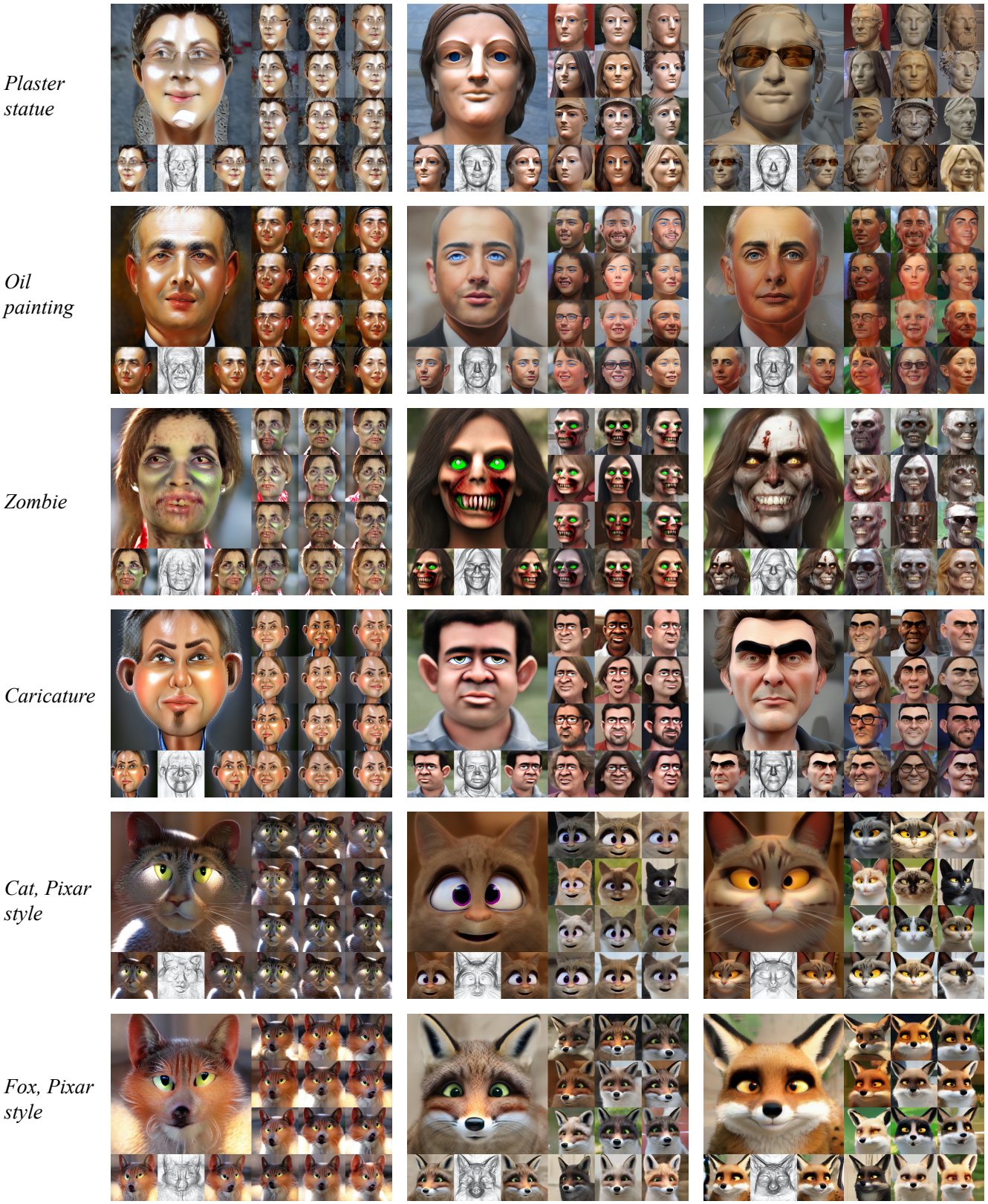
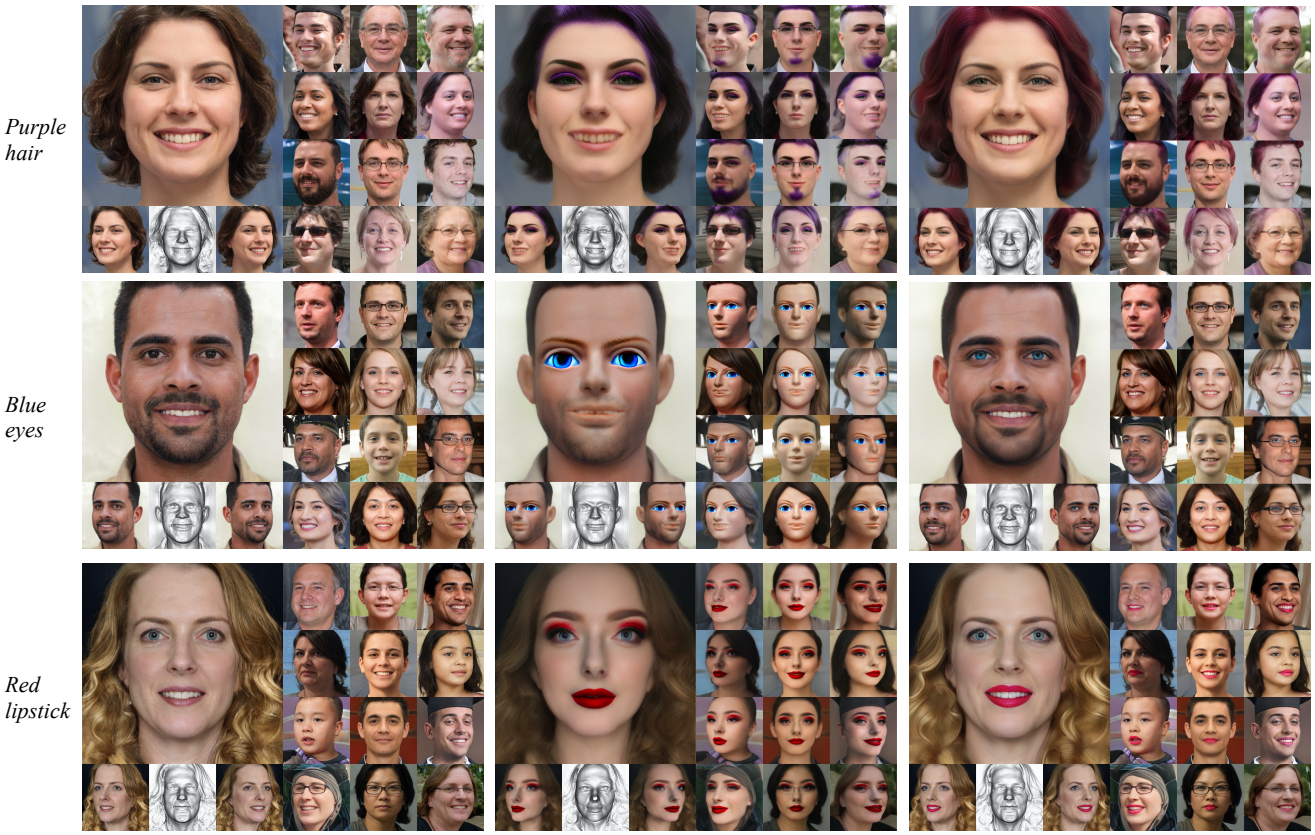


Figure 1. The qualitative comparisons for 3D domain adaptation on EG3D-FFHQ [2] (the former four rows) and EG3D-AFHQ (the last two rows). (Zoom in for a better view.)



(a) EG3D-FFHQ (source domain) (b) StyleGAN-Fusion (c) Ours

Figure 2. The qualitative comparisons for 3D-aware local editing on EG3D-FFHQ [2]. (Zoom in for a better view.)

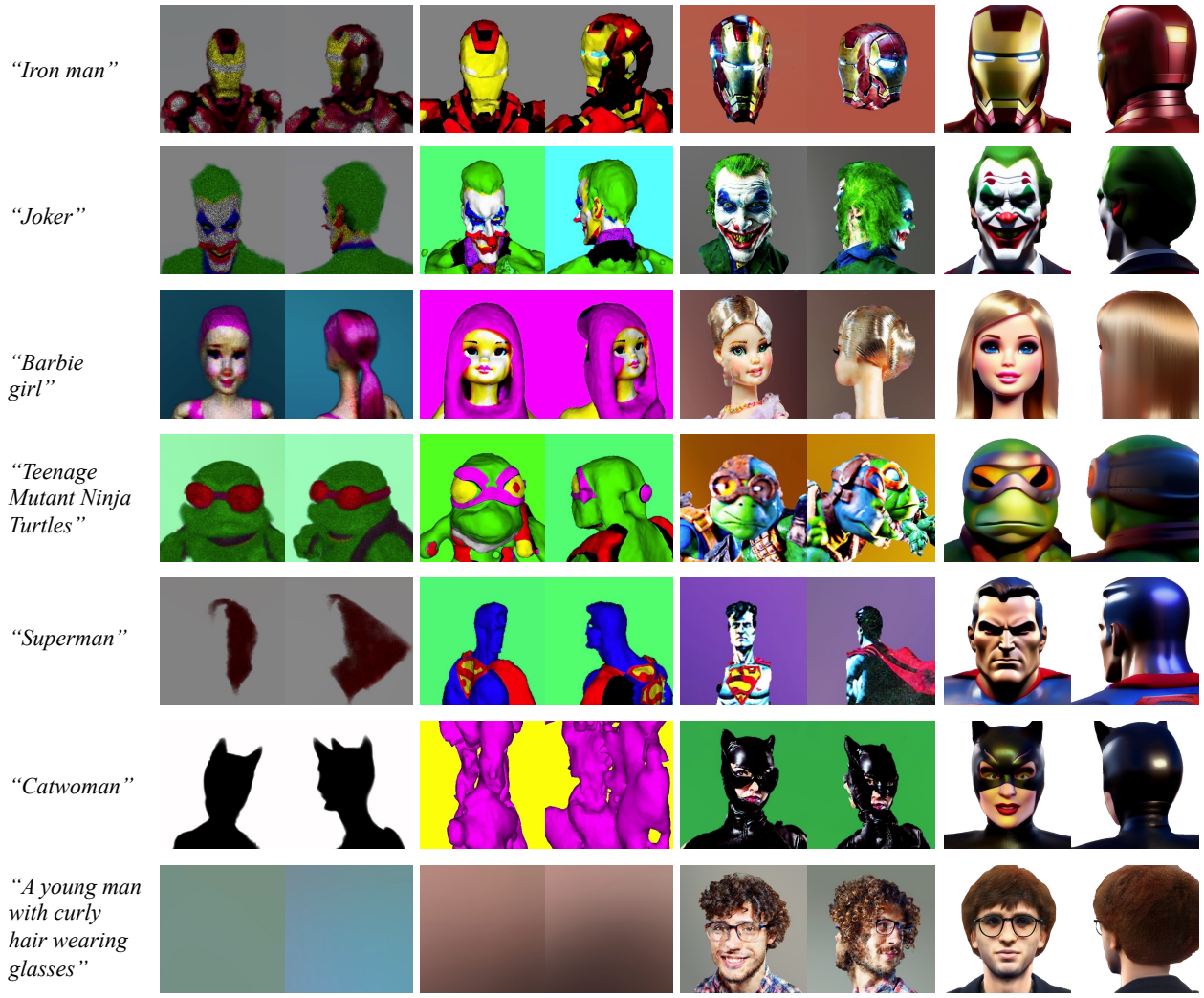


Figure 3. Visual comparisons on text-to-avatar task. The upper parts are the results of "head" and the lower parts are the results of "body". The empty images indicate convergence failure. (Zoom in for a better view.)

Figure 4. The geometry and multi-view rendering results of the first stage of our method given the text “catwoman”.

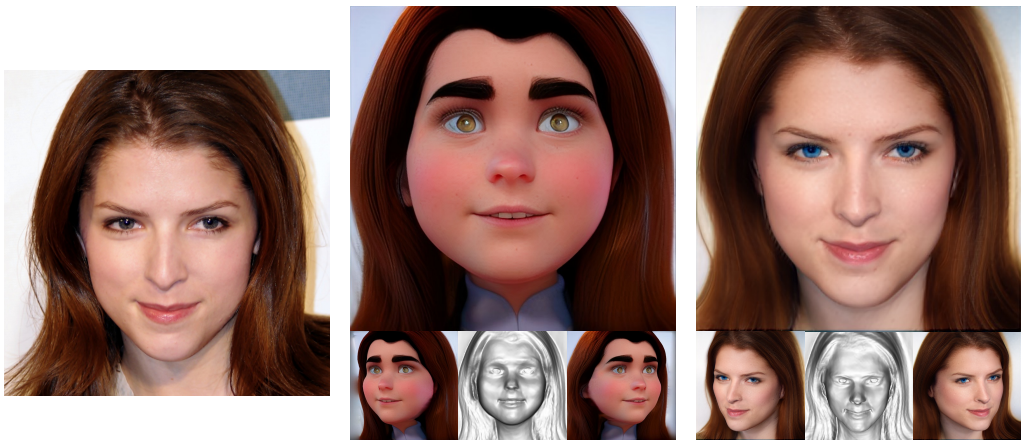


Figure 5. An example of the application of our method on 3D-aware stylization and local editing.