Supplementary Material ID-to-3D: Expressive ID-guided 3D Heads via SDS

Francesca Babiloni, Alexandros Lattas, Jiankang Deng, Stefanos Zafeiriou
Imperial College London, UK
https://idto3d.github.io
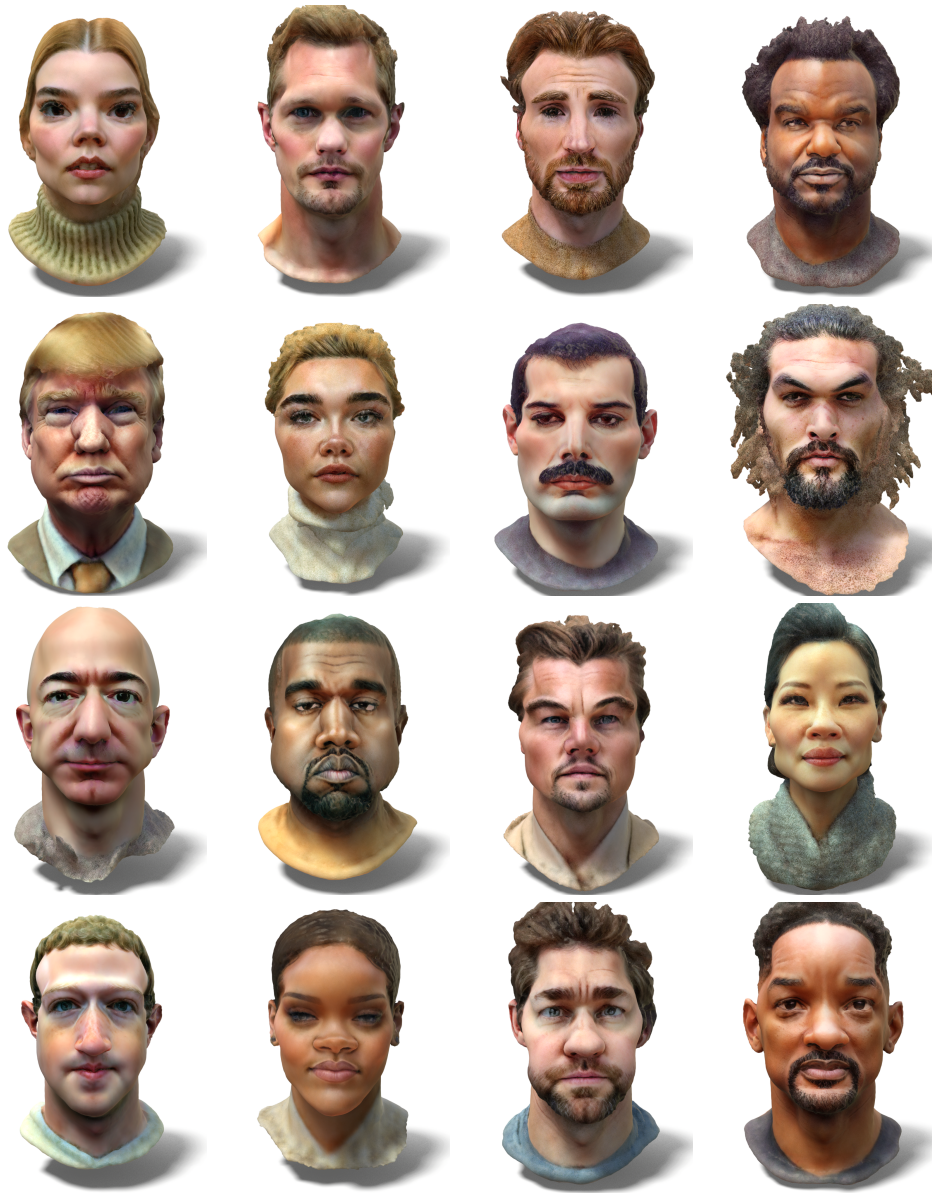
Figure 1: **3D assets** created by ID-to-3D can be photorealistically rendered in common 3D engines.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Figure 2: **ID-consistent texture and geometry generation**. The input images for ArcFace conditioning (**Left**) have been created using Stable-Diffusion. Normal maps (**Center**) are displayed next to renderings in studio lighting (**Right**). ID-to-3D creates ID-consistent geometry and textures with precise alignment.

Figure 3: **Randomized lighting** (**Left**) creates sharper textures and realist colors. Results generated by training with fixed lighting (**Right**) struggle to create crisp details.

1 Additional Results

We provide additional results for our methods. Figure 1 displays neutral expression for additional subjects, showcasing realistic, high-detail, ID-conditioned 3D assets that can be relit under various conditions. Figure 10 reports additional comparisons with text-to-3D SDS pipelines specialized in the generation of human avatars via SDS, while Figure 11 displays additional comparisons with methods that leverage text and images to create 3D assets. All methods are presented using the same rendering conditions. As apparent from the figures, the comparisons are consistent with the results presented in the main manuscript. ID-to-3D achieves the highest degree of geometric and texture quality. Figure 2 present ID-conditioned examples for AI-generated identities created starting from the test cases of the NHPM dataset, demonstrating results consistent with the findings presented in the main manuscript.

Visualization in Video Format. Taking in input only unconstrained pictures of a subject, ID-to-3D produces high-fidelity shapes and textures that can be photorealistically relighted in an arbitrary environment and illumination. Moreover, ID-to-3D establishes a new State-of-the-Art in the generation of 3D consistent human heads. Relighting videos for ID-to-3D's 3D heads assets that are shown in this paper, can be explored using our project page, together with video comparisons of existing SDS methods under fixed rendering conditions.

Impact of Randomized Lighting. We provide an ablation study on the impact of randomized lighting during texture training. As apparent in Figure 3 the use of this augmentation, paired with an albedo-oriented 2D guidance, allows for the generation of crisp details and vibrant colors, not achievable with standard training.

Visualization of 2D Guidance Images. Figures 5 and 6 display images generated by our 2D geometry-oriented and albedo-oriented guidance models for a range of identities, poses, and expressions. As visible, after convergence, our models are able to convincingly separate geometric and texture information for a given subject, realistically portrait side and back views, and reliably convey a wide range of id-consistent expressions. In Figure 7, we provide comparisons of our specialized 2D models against two well-established text-to-image models traditionally used in SDS pipelines: Stable-Diffusion v2.1 [3], DreamLike-PhotoReal [1]. As visible, traditional Stable-Diffusion models have three main shortcomings that limit their applicability to the task at hand: (1) Low ID retention: the models struggle to consistently create outputs of specific identities since they rely only on textual

2
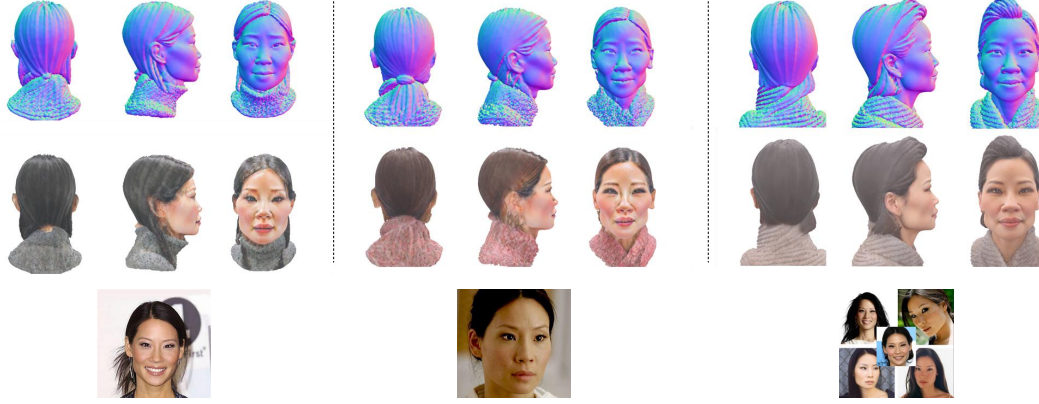
Figure 4: **Impact of Different ID-embeddings.** All the 3D heads are generated using a Neutral expression and rendered under the same lighting conditions. The first row displays the images used as input. Geometry is displayed as normal maps in camera coordinates.

prompts. (2) Low expressivity: the use of natural language to enforce expression conditioning is ineffective. The models overlook expression-related prompts to boost photorealism. (3) Inconsistent lighting: the models generate a wide range of lighting conditions, to enhance photorealism and artistic effect. This complicates the separation of lighting and albedo contributions when creating ready-to-render assets with SDS. In contrast, our 2D generators effectively separate appearance and geometry information while ensuring high expressivity, making them suitable for creating ID-driven 3D heads with expression control.

Impact of Different Identity Conditionings. Our method uses identity embedding to guide the geometry and texture generation of facial details, and is therefore bounded by the ability of these representations to convey in a concise and distinct manner the unique features of each face identity. Figure 4 compares the results of our final pipeline with that using less robust identity conditioning. We compare a 3D asset obtained using the identity embeddings derived from five images, with two different 3D heads created using a single in-the-wild picture of the same subject. Our pipeline is capable of creating 3D consistent heads in all the considered cases. However, The ArcFace features collected from a small poll of images provide better identity conditioning, with fewer texture artifacts and closer ID similarity with the reference identity.

2 Implementation Details

Implementation Details: 2D Guidance. We develop the geometry-oriented and albedo-oriented 2D guidance models by finetuning a large-scale diffusion model on a small dataset of real 3D scans. This representation serves as identity conditioning for the geometry-oriented and albedo-oriented 2D diffusion models. We follow [6] and use normal maps in camera coordinates as a 2D proxy for geometric information. To encourage the separation of color and lighting information, we select albedo maps as texture data. We modified a large-scale Stable Diffusion model pre-trained to perform photorealistic 2D face portrait generation [8], and finetuned the geometry and albedo model in two stages to minimize identity drifting. First, we modified the architecture with LoRA layers to perform a style transfer task. Then, we modified and trained the same architecture to accommodate expression conditioning. Both fine-tuning is done for $200k$ iterations with early stopping with a lr of $1e-4$ on 8 Nvidia-V100 GPUS. We use 16 as the hidden LoRA size for self-attention layers. We use LoRA for cross-attention layers with a hidden size 32. For each image, we extract the ArcFace identity embedding after cropping and centering. The identity embeddings are then concatenated and processed by a shallow 2-layer MLP to match the dimension of the text features in the pretrained diffusion model. For training, we use the recently released neural head parameter data set [5], comprising 255 subjects and 23 facial expressions. We selected 25 subject to test the drifting ID during training. For the geometry-oriented 2D guidance model, we created a training set of $250k$ samples that extract 2D normal maps in camera coordinates from random subjects in random expressions and camera poses. For the texture-oriented 2D guidance model, we extracted a dataset of $250k$ samples from the texture of the objects. We extracted text descriptions for each render and supplemented the training with a text prompt containing camera pose, gender, ethnicity,

3

Figure 5: **Camera aware 2D generation** enables the generation of samples aligned with different camera poses. We display images for back views (**Left**), side views (**Center**) and front views (**Right**).

Figure 6: **Generation of id and expression conditioned images.** The geometry-oriented model (**Right**), creates high quality 2D normal maps that can be used as proxy for geometric information. The albedo-oriented 2D model (**Left**) generate images with consistent lighting that can be used as guidance for high-quality texture generation.

Figure 7: **Comparison of image generation with id and expression conditioning.** Images generated by our albedo-oriented model (**Row 2**), our geometry-oriented model (**Row 3**), Stable-Diffusion v2.1 model (**Row 4**), DreamLike-Photoreal model (**Row 5**). All image generators are prompted using a common textual prompt *"A DSLR face portrait of ..."* and the same id/expression specific text (**Row 1**). Traditional text-to-2D models struggle to create consistent lighting and expressive faces.

4

and age group for each subject (e.g., 'A front view normal map face portrait of a young Caucasian female on a white background').

Implementation Details: Geometry and Texture Generation. Given an ArcFace identity, geometry generation is achieved through a two-stage pipeline. First, we follow [4] and initialize the 3D geometry with a FLAME head template by regression fitting. Then, we train the geometry model using randomly selected camera poses and expression conditioning. We used a neural head expression model as a hybrid 3D representation for geometry. We used HashGrid [7] encoding and a grid with 256 resolution. We used a 4 layers transformer and jointly trained 3, 6, or 13 expressions each associated with its unique latent code of dimensions 32. Our model can train a neutral identity representation in 6000 iterations, which takes approximately 30 minutes on v100 GPUS. Our model can jointly train a larger set of 13 expressions in approximately 4 hours. We use a learning rate of $1e-4$. We used a random sampling strategy and considered time steps in the range $[200, 700]$ for the first half of the training. We use an annealing strategy with time steps in the range $[200, 50]$ in the second half of the training. We use a score distillation sampling loss with a guidance weight 10 and a Laplacian smoothing loss with a weight 5000.

Given a learned geometry model, the texture generation is achieved via a three-stage pipeline conditioned on the same ArcFace identity. We train the texture model using randomly selected camera poses and expression conditioning. We use random lighting during training, using a list of 60 HDR maps and a set of random augmentations. We trained a 3 layers transformer and jointly trained 3, 6, or 13 expressions each associated with its unique latent code of dimensions 32. Our model can train a neutral identity representation in 2000 iterations, taking approximately 15 minutes on v100 GPUS. Our model can jointly train a larger set of 13 expressions in approximately 10000 iterations. As first step, the model is optimized to learn only the diffuse term of the textures. This portion of the training uses 80% of all available iterations, a learning rate of 0.01, and time steps in the range $[50, 900]$. The second step then proceeds to jointly optimize the roughness and metallic term of the texture together with the diffuse term, using the same learning rate but a different timesteps range $[50, 500]$. Lastly, the pipeline uses an optional and quick refinement stage with the goal of introducing high-frequency details. The last 20 iterations use a large photorealistic text-to-image model as guidance and deploy timesteps in the range $[50, 100]$. All stages use a score distillation sampling loss with weight 10.

Implementation Details: User Study. In the main paper, we evaluate the 3D heads generated by ID-to-3D in terms of perceptual geometric quality and texture quality through a user study. We follow [6] and compare with four state-of-the-art methods on 30 prompts. We collect the preferences of 50 participants using anonymous online survey forms. To fairly compare among methods, we created 3D assets using comparable input (i.e. same text and image prompt), and gathered as visualization a GIF accumulating 360 renderings under the same lighting and camera conditions. To assess geometric quality, we remove the contribution of texture by visualizing normal maps. Each volunteer was asked to: 1) Choose which 3D head among the 5 options represents the most appealing, detailed, and ready-to-use 3D object. 2) Choose which 3D head among the 5 options represents the most appealing, detailed, and ready-to-use 3D geometry. As visible in the main manuscript, ID-to-3Daccumulates higher preferences for both metrics, showing superior performance compared to both human-specific text-based methods and image-to-3D baselines.

Results Rendering. For the rendering of results shown in this paper, as well as in the supplemental video files, an off-the-shelf commercial rendering engine has been used [2], highlighting the ease of integration of our results to the tools of the industry. In that manner, any similar standard rendering engine could also be used.

3 Additional Analyses

Dataset Statistics. We train our model using the NPHM dataset, which is rich in expression data and geometric quality. To use personalized text prompt during training, we semi-automatically annotate the provided assets as a pre-processing step. An overview of the statistics of the dataset is shown in Figure 8 (Right). During the 2D guidance training, we use gender, age, ethnicity, and hairstyle as textual prompts to guide generation. As a result, our method generates 3D heads with diverse identities, ethnicities, and ages. However, we acknowledge the imbalanced nature of the dataset, which could result in the underrepresentation of minorities.

5
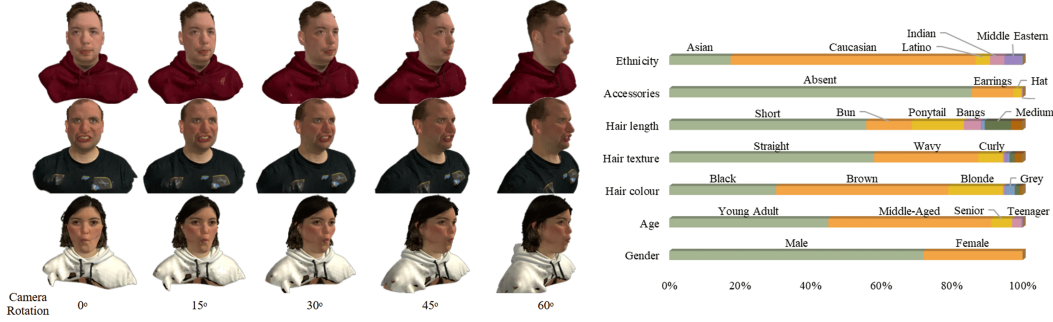
Figure 8: **Dataset Visualizations. (Left)** Renders at various camera angles for different subjects of the NPHM dataset. **(Right)** NPHM dataset statistics for gender, age, ethnicity, and hairstyle choices.
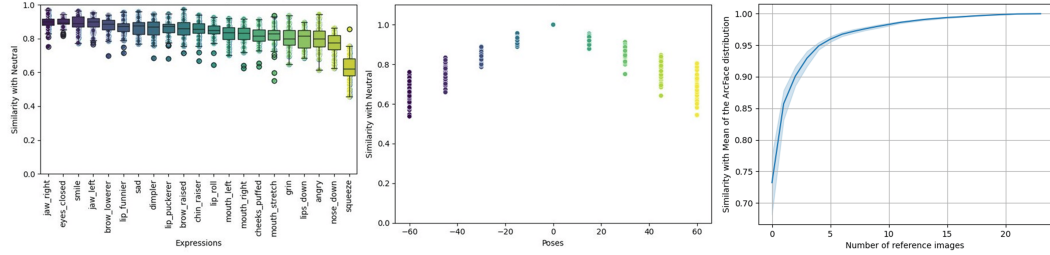
Figure 9: **Analyses on identity embeddings. (Left)** Cosine similarity between ArcFace of neutral expression and remaining expressions for each subject in the training dataset. **(Middle)** Cosine similarity between ArcFace of frontal pose and 9 alternative poses captured with different camera angles for each subject in the training dataset. **(Right)** Relationship between identity similarity and number of reference images. The identity embeddings are robust to expression variety, camera choices, and number of images used.

Robustness of the ID Embeddings. We provide analyses on the robustness of ID embeddings with respect to the expressions and poses of the camera in Figure 9. We use the 3D assets provided in our training dataset. For each identity and expression, we collect renders for 9 different camera rotation angles [-60°, -45°, -30°, -15°, 0°, 15°, 30°, 45°, 60°] and extract their identity embedding (ArcFace). Examples of renders are visible in Figure 8 (Left). We use a neutral pose with a 0° rotation angle as a reference. Note that, according to definition, a similarity score above 0.5 signifies the same identity. The impact of the expressions on the ID embeddings (Fig. 9 (Left)) is isolated by computing the cosine similarity between the neutral reference and all remaining expressions captured with a rotation angle of 0°. The impact of the camera pose on ID embeddings (Fig. 9 (Middle)) is isolated by computing the cosine similarity between the neutral reference and the neutral expression captured with all the 9 possible rotation angles. As clearly visible, ArcFace reliably captures identity features across various expressions and poses, showcasing robust behavior even for extreme expressions (e.g. "Squeeze" Avg-SimID: 0.62) and substantial camera rotation (e.g. -60° Avg-SimID: 0.7).

Relationship between ID similarity and number of reference images. We provide analyses on the robustness of ID embeddings with respect to the number of reference images used to capture the identity of a subject in Figure 9 (Right). We consider 40 subjects, 25 in-the-wild images for each subject, and extract for each image its identity embedding (ArcFace). For each subject, we consider the center of the distribution as representative of its facial features. We report the similarity between the center of the distribution and the mean ArcFace created with a subset of N number of reference images. The plot shows the averaged trend for 40 identities (blue line) together with its standard deviation (light blue). The trend reaches a plateau after 20 images, while 5 images is enough to reach an identity similarity of more than 0.95 for all the IDs considered. In our experiments, we selected 20 images to use as references for our comparisons and kept 5 images to use as input to our method, ensuring a good trade-off between identity similarity retention and practicality.

6
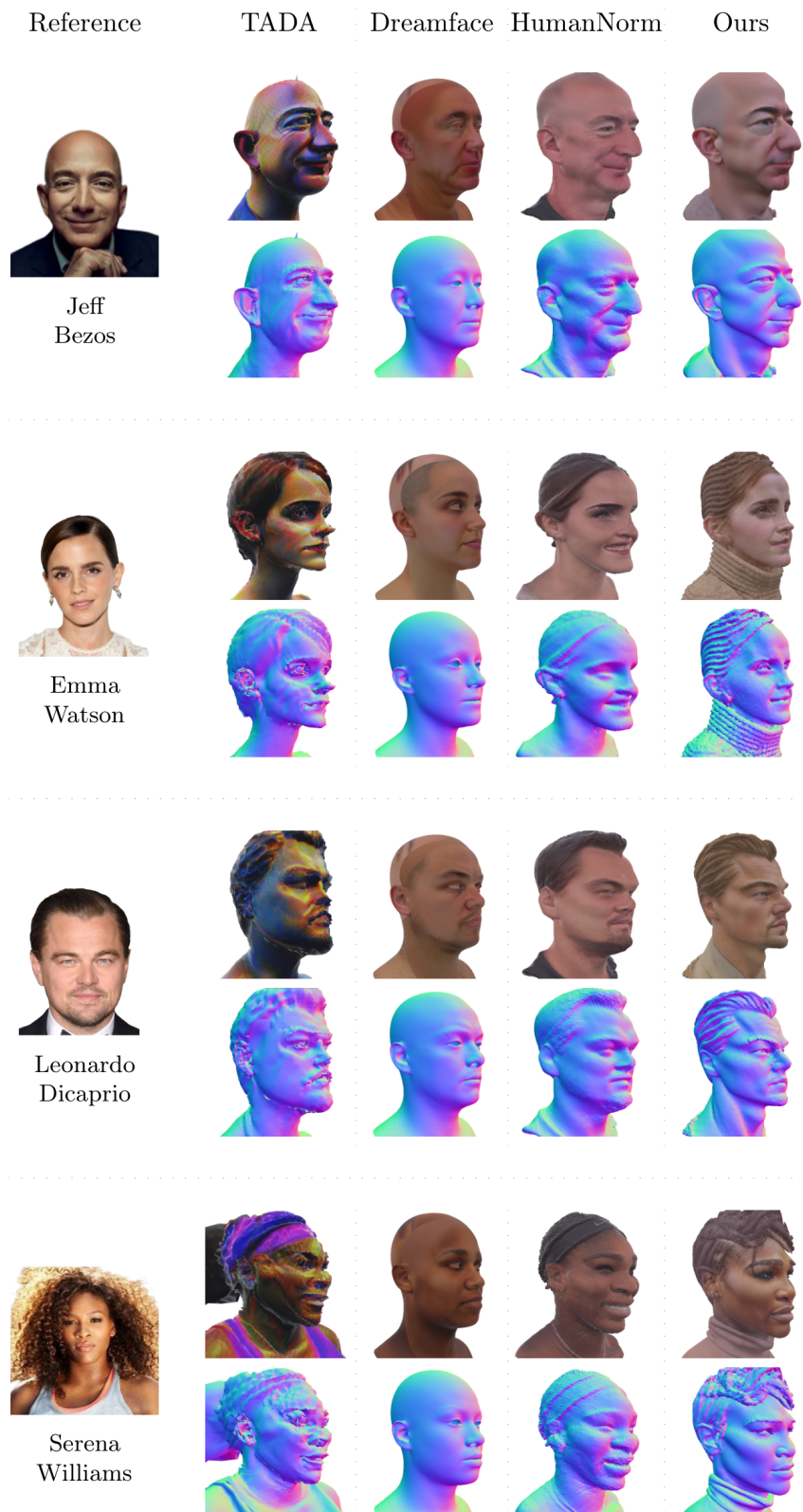
Figure 10: **Comparisons with text-to-3D generation methods.** ID-to-3D creates realistic 3D heads when compared with human-centric SDS methods based on text prompts.

7
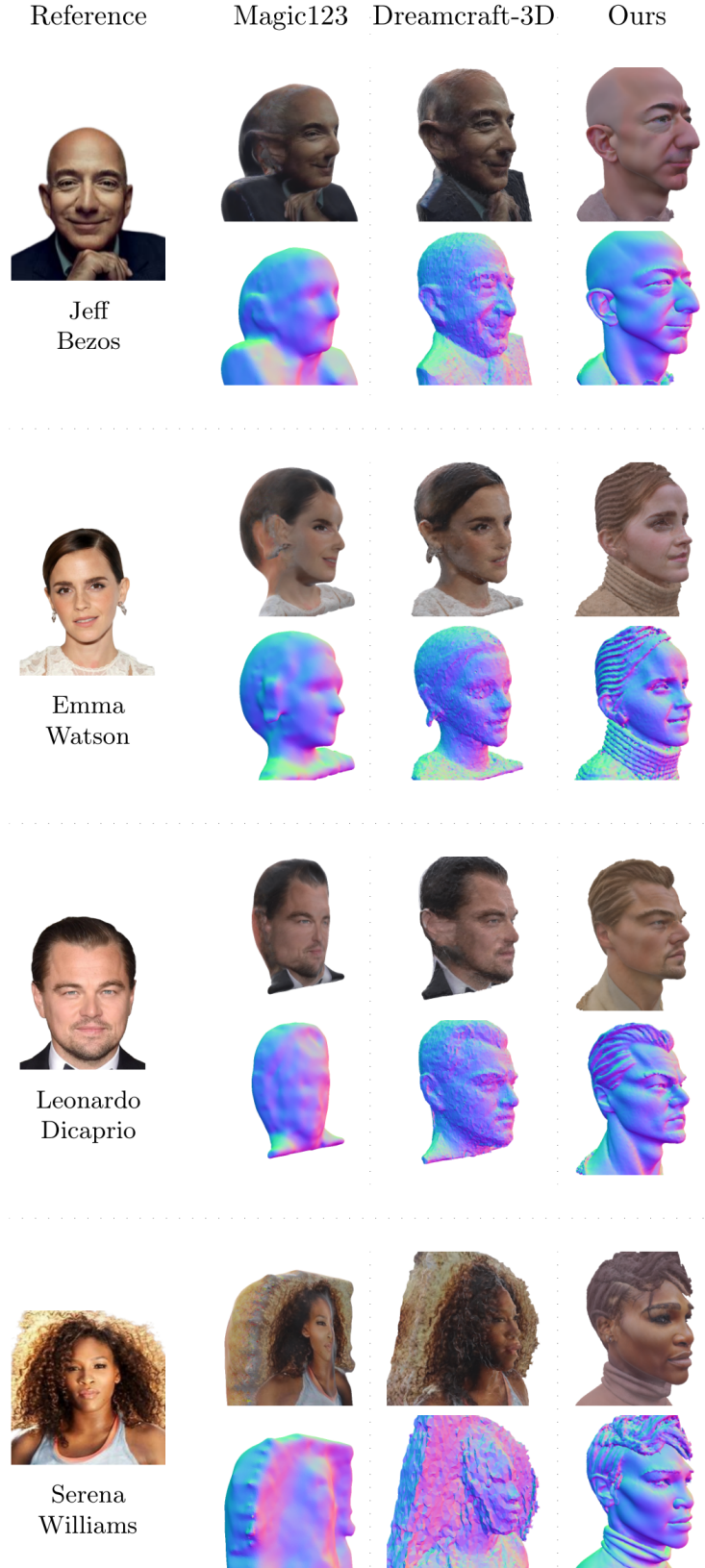
Figure 11: **Evaluating text+image-to-3D generation techniques.** ID-to-3D consistently produces high-quality 3D heads across various viewpoints, outperforming state-of-the-art SDS methods utilizing text prompts and images.

8

References

[1] Dreamlike-photoreal 2.0. `https://huggingface.co/dreamlike-art/dreamlike-photoreal-2.0`.

[2] Marmoset toolbag. `https://marmoset.co/toolbag/`. Accessed: 2022-05-20.

[3] Stable diffusion 2.1. `https://huggingface.co/stabilityai/stable-diffusion-2-1`.

[4] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22246–22256, 2023.

[5] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Learning neural parametric head models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21003–21012, 2023.

[6] Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, and Qing Wang. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. *arXiv preprint arXiv:2310.01406*, 2023.

[7] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022.

[8] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

9