SpinMeRound: Consistent Multi-View Identity Generation Using Diffusion Models

Supplementary Material

Stathis Galanakis[1] Alexandros Lattas[1] Stylianos Moschoglou[1]
Bernhard Kainz[1,2] Stefanos Zafeiriou[1]

[1]Imperial College London, UK
[2]FAU Erlangen–Nürnberg, Germany

1. Limitations and Future Work

Although SpinMeRound showcases high-fidelity results, it has some limitations. More specifically, our model takes over structural limitations presented in Panohead [1] due to the synthetic dataset used. This means that there are some inconsistencies in the generated eyes, hair and noses. Moreover, using the alignment pipeline sometimes results in failure cases, due to misalignment.

All in all, the fact that we do not use any captured data limits our model's capabilities. Hence, this is a direction that we plan to explore in future work. Captured datasets such as FaceScape [13], Renderme-360 [8] and NeRSemble [7] can be used to further improve our results. Finally, integrating video diffusion models [2] can be another direction for our future work, to improve the consistency of the generated viewpoints.

2. Training Details

SpinMeRound begins training using the publicly available Arc2Face model [10]. The Arc2Face model is built upon *Stable Diffusion 1.5* [11], meaning that it incorporates the following preconditioning functions, according to the EDM framework [6]:

$$c_{skip}^{SD1.5}(\sigma) = 1, \qquad c_{out}^{SD1.5}(\sigma) = -\sigma,$$
$$c_{in}^{SD1.5} = \frac{1}{\sqrt{\sigma^2+1}}, \quad c_{noise}^{SD1.5}(\sigma) = \underset{j\in[1000]}{\arg\max}(\sigma - \sigma_j)$$

As proposed in [6], we modify the aforementioned preconditioning by:

$$c_{skip}(\sigma) = (\sigma^2+1), \qquad c_{out}(\sigma) = \frac{-\sigma}{\sqrt{\sigma^2+1}},$$
$$c_{in} = \frac{1}{\sqrt{\sigma^2+1}}, \qquad c_{noise}(\sigma) = 0.25\log\sigma,$$

Furthermore, we use the proposed noise distribution and weighting functions $log\sigma \sim \mathcal{N}(P_{mean}, P_{std}^2)$ and $\lambda(\sigma) = (1+\sigma^2)\sigma^{-2}$, with $P_{mean} = 0.7$ and $P_{std} = 1.6$. We finetune the pre-trained Arc2Face model for 31k iterations, using the training dataset provided by the Arc2Face authors.

2.1. Shape Normals Retrieving

Algorithm 1 Shape Normals sampling using Guidance

Input: The aligned facial "in-the-wild" image $\bar{\mathbf{I}}$, the gradient scale α, the binary visibility mask m, the conditioning mechanism $\mathcal{C}$, and encoder $\mathcal{E}$.

$\mathbf{c} \leftarrow \mathcal{C}(\bar{\mathbf{I}})$, $\mathbf{z}_{gt} \leftarrow \{\mathcal{E}(\bar{\mathbf{I}})|\mathbf{0}\}$
$\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, t_0^2\mathbf{I})$
for all i **from** 0 **to** N-1 **do**
  $\epsilon_i \sim \mathcal{N}(\mathbf{0}, S_{noise}^2\mathbf{I})$
  $\gamma_i = \begin{cases} \min(\frac{S_{churn}}{N}, \sqrt{2}-1) & \text{if } t_i \in [S_{tmin}, S_{tmax}] \\ 0 & \text{otherwise} \end{cases}$
  $\hat{t}_i \leftarrow t_i + \gamma_i t_i$
  $\hat{\mathbf{x}}_i \leftarrow \mathbf{x}_i + \sqrt{\hat{t}_i^2 - t_i^2}\epsilon$
  $\mathcal{L} \leftarrow ||(\mathbf{z}_{gt} - D_\theta(\hat{\mathbf{x}}_i; \hat{t}_i, \mathbf{c})) \odot m||_2^2$
  $\mathbf{d}_i \leftarrow (\hat{\mathbf{x}}_i - D_\theta(\hat{\mathbf{x}}_i; \hat{t}_i, \mathbf{c}) - \alpha\frac{\partial\mathcal{L}}{\partial\hat{\mathbf{x}}_i})/\hat{t}_i$
  $\mathbf{x}_{i+1} \leftarrow \hat{\mathbf{x}}_i + (t_{i+1} - \hat{t}_i)\mathbf{d}_i$
end for
return $\mathbf{z}_N$

As mentioned in Section 3.3, given an input "in-the-wild" facial image, we first extract the respective shape normals $\mathcal{N}$. Our proposed sampling methodology is presented in Algorithm 1 and is inspired from Relightify [9]. Given an aligned "in-the-wild" image, we follow the sampling algorithm presented in Algorithm 1, where $\odot$ denotes the Hadamard product $\bar{\mathbf{I}}$. We guide the sampling process to generate the respective shape normals, based on the distribution of the training data. In detail, we firstly extract the conditioning label, as described in Section 3.1, and the latent feature maps of the image $\bar{\mathbf{I}}$, which gets padded. After

Figure 1. We compare the generated backhead between SpinMeRound(Ours) and Panohead [1].

Figure 2. SpinMeRound and Eg3D [3] are shown at $+90^{\circ}$ angle.

that, we sample the input gaussian noise. For each sampling step, we estimate $\hat{x}_i$ as presented in steps 4, 5, 6 and 7. Then, we compute the guidance loss by calculating the masked *L2*-distance between the ground-truth latent vector $\mathbf{z}_{gt}$ and the estimated $D_\theta(\hat{\mathbf{x}}_i; \hat{t}_i, \mathbf{c})$. We calculate the Euler step from $\hat{t}_i$ to t_{i+1} by applying the formula in line 9. During sampling we set the guidance scale equal with 10^5 and we run for $t = 50$ sampling steps. The sampling process takes about 2.4 seconds while it runs on an NVIDIA A100-PCIE.

3. Qualitative comparison with Panohead and Eg3D

Panohead [1] is a NeRF-based method capable of generating 360° views. Given an input facial image, it requires a fitting process to produce novel views, often necessitating additional pivotal tuning. In contrast, SpinMeRound eliminates the need for any fitting or fine-tuning steps. Additionally, as presented in Fig. 1, Panohead frequently introduces artifacts on the back of the head, a limitation our method overcomes. On the other hand, EG3D [3] is another NeRF-based method having similar drawbacks as Panohead. Moreover, it only focuses on generating near-frontal views contrary to our full-head approach as shown in Fig 2. We present a qualitative comparison between our method and Panohead [1] in Fig. 4. As demonstrated, our proposed approach achieves more faithful novel view synthesis of the input subjects and significantly reduces the back-of-head artifacts commonly observed in Panohead.

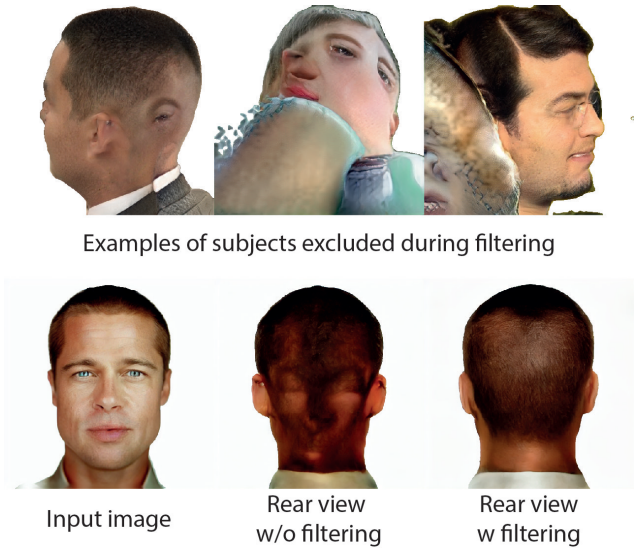

Figure 3. Top row: Examples of filtered-out subjects. Bottom row: We present the rear view of an input image (left) when given to a model trained with the unfiltered and the filtered datasets.

Importance of the used dataset To train our method, we utilize a synthetic dataset, acquired using Panohead [1]. Since artifacts frequently occur during sampling, a manual filtering step is necessary. We filtered out failure cases, including artifacts or abnormal side-heads, that could negatively impact the performance of our model. We present excluded samples in the first row of Figure 3. To confirm this, we trained two models for 100 epochs: (a) one using the prior-filtering dataset, and (b) one using the filtered dataset. Then, we gathered 50 images and compared their identity similarity scores at a yaw angle of 45°. The first model achieved an identity similarity score of 0.651, whilst the model trained on the filtered dataset achieved a score of 0.72. Moreover, we showcase the generated real views of a representative subject in the bottom row of Figure 3.

4. Identity sampling

As mentioned in Section 5.1 and presented in Figure 8 of the main paper, our method can generate multi-view human identities, given only the input embedding. Although, this work does not focus on multi-view identity sampling, we explore our method's capabilities in this section.

As SpinMeRound has been trained using the classifier-free guidance (CFG)[5] whilst getting 0, 1 or 3 conditioning input images, it can be used to conditionally generate novel images depicting a similar identity as the input one. By setting the guidance scale equal with 3.5, we run the EDM sampler [6] for 50 sampling steps. We set

$S_{churn} = 0, S_{tmin} = 0.05, S_{tmax} = 50, S_{noise} = 1.003$ and we use the EDM [6] discretization steps, with maximum sigma equal to 700. The sampling process takes about 10 seconds while it runs on an NVIDIA A100-PCIE. We present samples generated from our model in Figure 8.

5. More samples

We provide additional results in Figures 5, 6, 7, 9 and 10. SpinMeRound is demonstrated under extreme viewpoints in Figure 5 whilst Figure 6 presents a qualitative comparison between our method's performance and Panohead [1], Eg3D [3], DiffPortrait3D [4] and SV3D [12]. In Figure 7, we showcase samples generated while using SpinMeRound, under $\{\pm 9^\circ, \pm 16^\circ, \pm 23^\circ\}$ elevation and azimuth angles. Additionally, samples produced from our model are presented in Figures 9 and 10, given the input images on the left. As illustrated, our proposed methodology can be applied to a wide variety of images, including diverse identities, input angles and image styles.

References

[1] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360°, 2023. 1, 2, 3, 4, 5

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 1

[3] 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, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16123–16133, 2022. 2, 3, 5

[4] 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 (CVPR)*, pages 10456–10465, 2024. 3, 5

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

[6] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Proc. NeurIPS*, 2022. 1, 2, 3

[7] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 42(4), 2023. 1, 5

[8] Dongwei Pan, Long Zhuo, Jingtan Piao, Huiwen Luo, Wei Cheng, Yuxin Wang, Siming Fan, Shengqi Liu, Lei Yang, Bo Dai, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, and Kwan-Yee Lin. Renderme-360: Large digital asset library and benchmark towards high-fidelity head avatars. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 1

[9] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, and Stefanos Zafeiriou. Relightify: Relightable 3d faces from a single image via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1

[10] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model for id-consistent human faces. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1

[11] 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. 1

[12] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion, 2024. 3, 5

[13] Hao Zhu, Haotian Yang, Longwei Guo, Yidi Zhang, Yanru Wang, Mingkai Huang, Menghua Wu, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: 3d facial dataset and benchmark for single-view 3d face reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023. 1

Figure 4. We present a qualitative comparison between SpinMeRound and Panohead [1] at yaw angles from $+45^{\circ}$ to $+315^{\circ}$. Contrary to Panodhead, our method faithfully generates novel views of the input subjects, without any artifacts on the backhead.

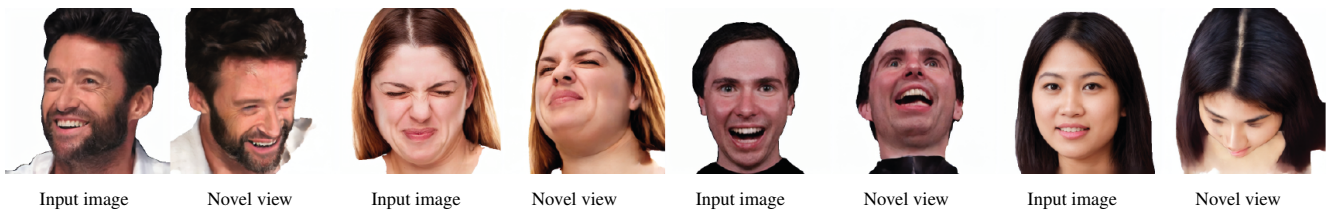

Figure 5. SpinMeRound is showcased under extreme angles.

Figure 6. Qualitative comparison on NeRSemble dataset [7]. As showcased, SpinMeRound shows superior results over Panohead [1], Eg3D [3], DiffPortrait3D [4] and SV3D [12].

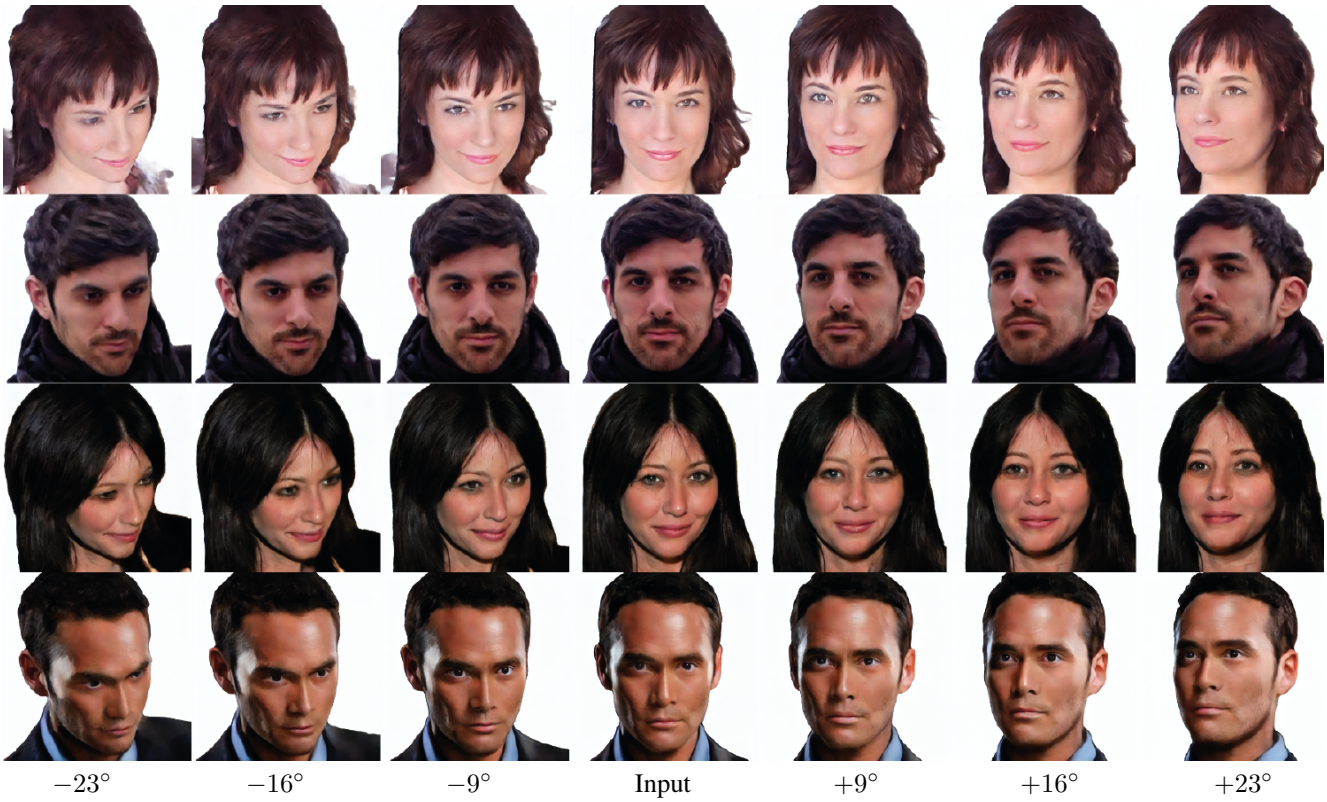

Figure 7. We showcase samples under $\{\pm 9°, \pm 16°, \pm 23°\}$ elevation and azimuth angles.

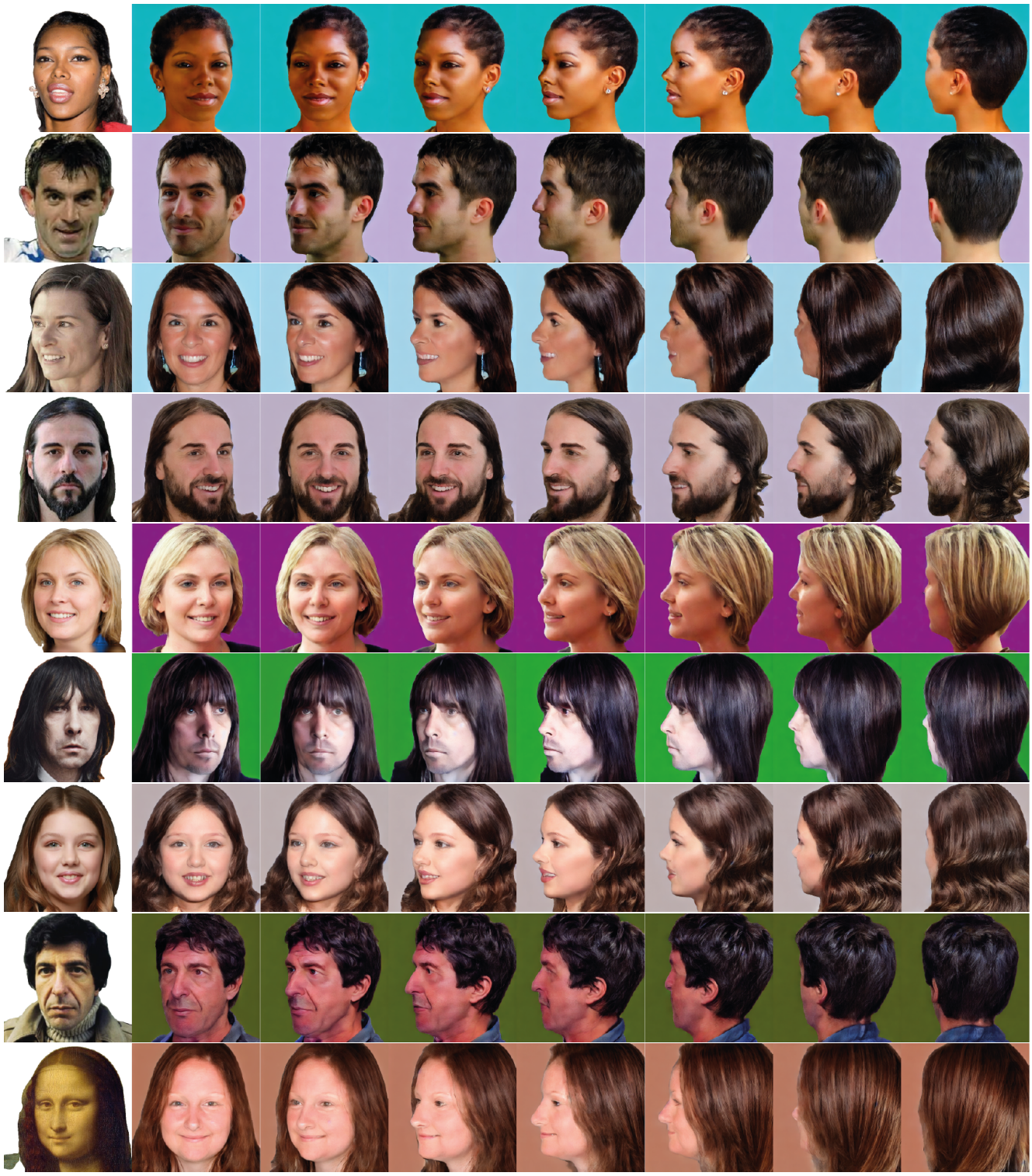

Figure 8. Samples generated using SpinMeRound using *only* the input identity vector.

Figure 9. Samples generated with our method, using the images on the left as input (1/2).

Figure 10. Samples generated with our method, using the images on the left as input (2/2).