Bridging the Gap: Studio-like Avatar Creation from a Monocular Phone Capture -Supplementary-

ShahRukh Athar[1,3*] Shunsuke Saito[2] Zhengyu Yang[2] Stanislav Pidhorskyi[2]
Chen Cao[2]

[1]Captions Research, New York [2]Meta Reality Labs, Pittsburgh.

[3]Stony Brook University, New York

shahrukh@nocapinc.com shunsuke.saito16@gmail.com
stpidhorskyi@meta.com zhengyu-yang@outlook.com
zju.caochen@gmail.com

1 Texture and Render Results

In Fig 1, we show more qualitative results of the texture maps generated with our method versus prior art. As can be seen, across all subjects, our method generates texture maps that have better quality illumination, facial details and inpainting of missing regions. In Figs 2-10, we show multiview renders of these subjects (and a few more) using the Universal Prior Model from AVA [1]. Again, we see that the avatars generated using the texture maps from our method are significantly more photorealistic than prior art.

2 Ablation of optimization resolution

Models	Full Network opt	Opt8 (Ours)	Opt16
FaceID ↓	$5.01e-4$	$4.31e-4$	$4.30e-4$
KID ↓	$1.36e-3$	$1.42e-3$	$1.63e-3$

Table 1: Ablation of optimization resolution. Best and Second Best scores are highlighted.

As described in Section 3.1 of the paper, we optimize $\mathcal{G}_{Studio}$ using the following loss:

$$\min_{\mathcal{G}^{\theta(8+)}_{Studio}} \max_{\mathcal{D}_{Studio}} \mathcal{L}_{Adv} + \mathcal{L}_{R1} + \mathcal{L}_{Percp-Recons} + \lambda_1 \mathcal{L}_{Percp} + \lambda_2 \mathcal{L}_{FaceID} \tag{1}$$

where $\mathcal{G}^{\theta(8+)}_{Studio}$ denotes optimizing the parameters of the generator after the 8×8 resolution. The intuition behind this is that identity-specific information is stored in the

[*] Work done while interning at Meta Reality Labs

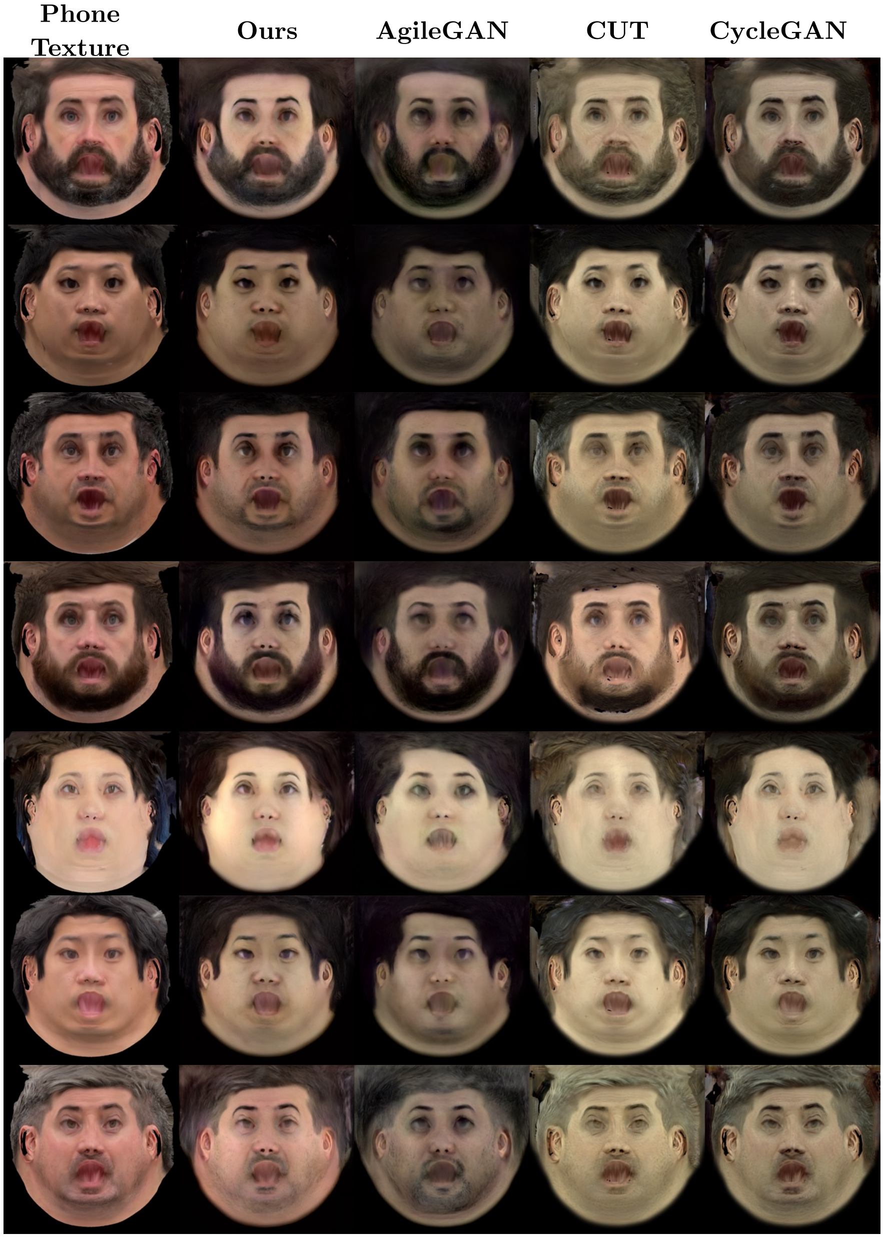

Fig. 1: Comparisons on Unpaired Phone-Captured Data: In comparison to prior work, our method excels in generating texture maps with superior preservation of identity, enhanced photorealism in facial details, more uniform illumination, and improved inpainting of missing regions.

low-resolution maps and freezing the parameters of the lower resolution boosts identity

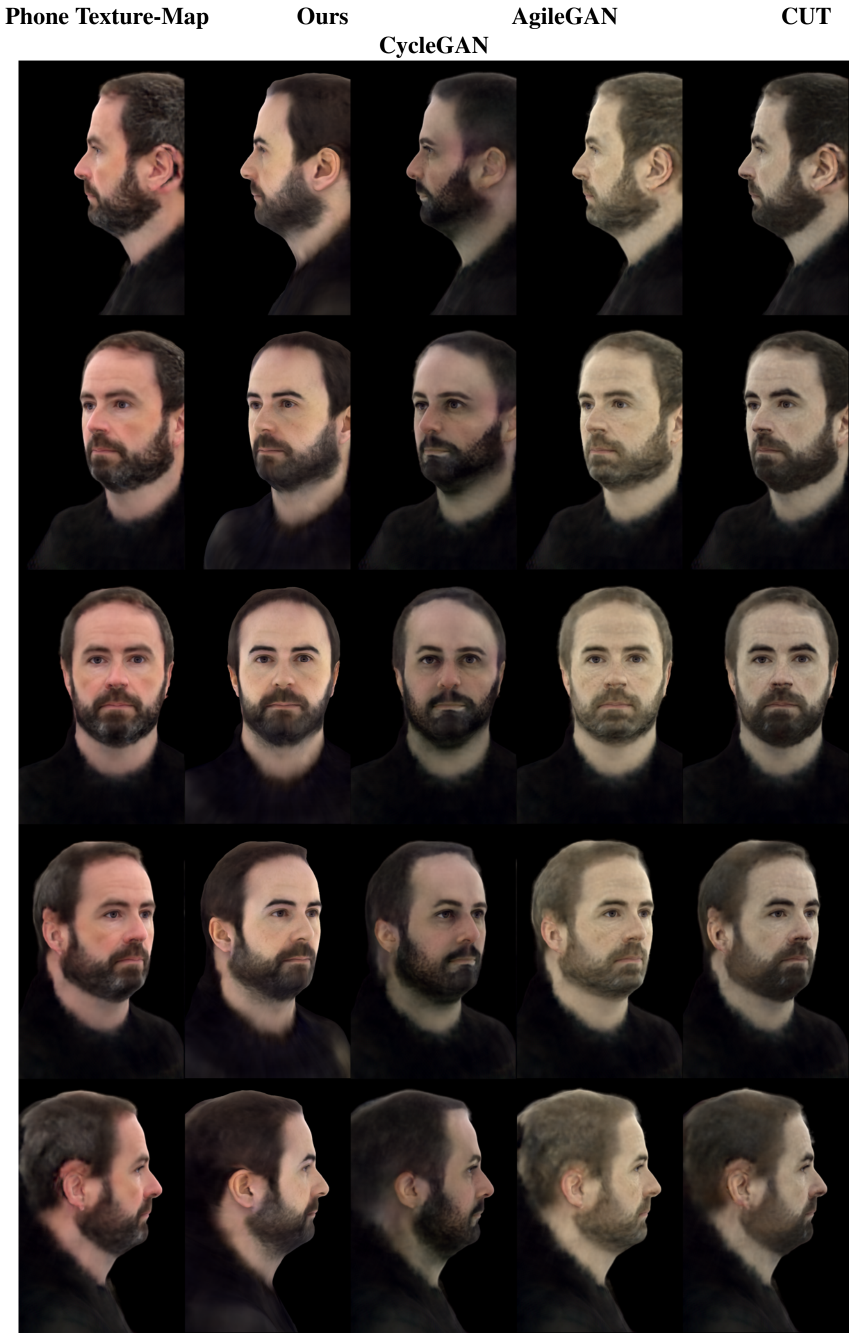

Fig. 2: Render using the Universal Prior Model from AVA [1]

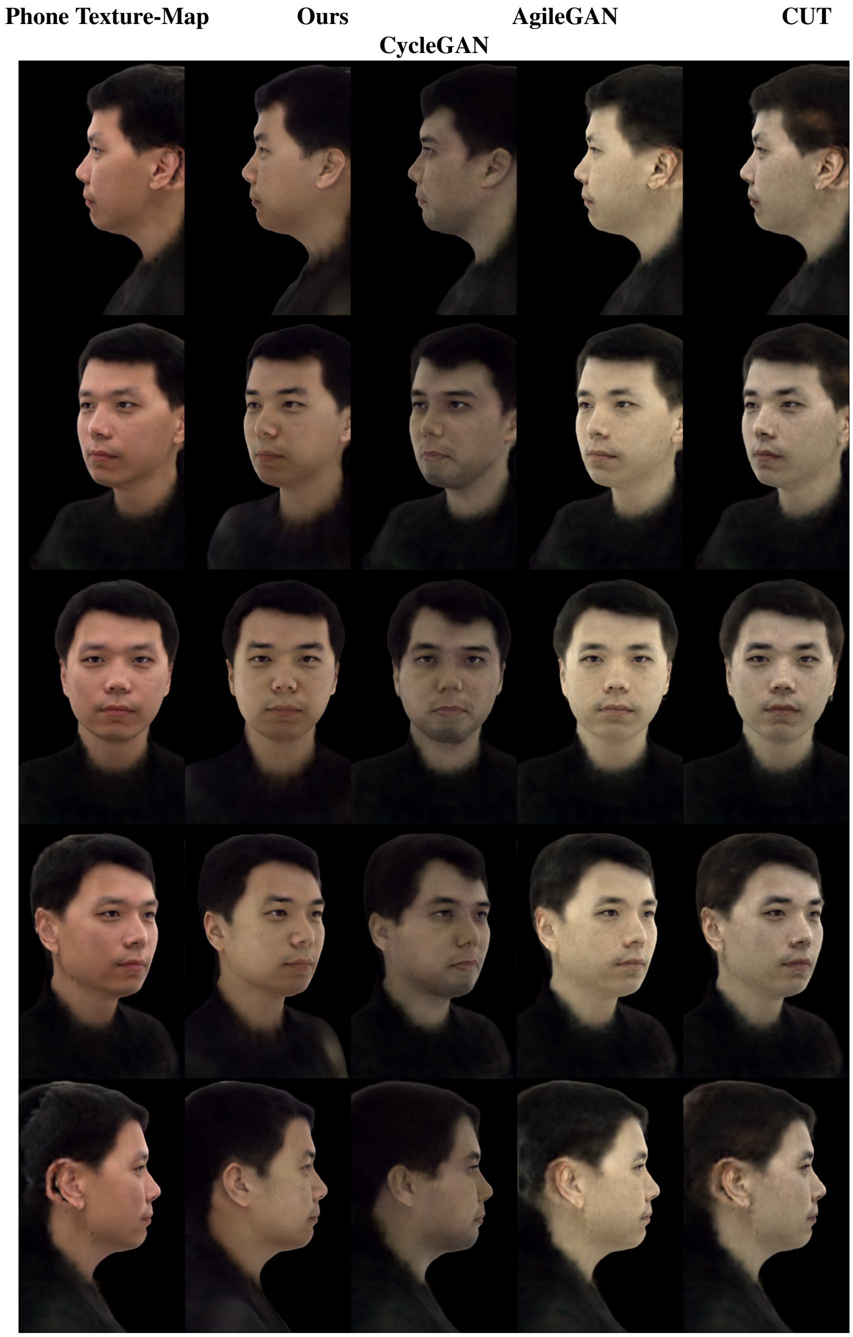

Fig. 3: Render using the Universal Prior Model from AVA [1]

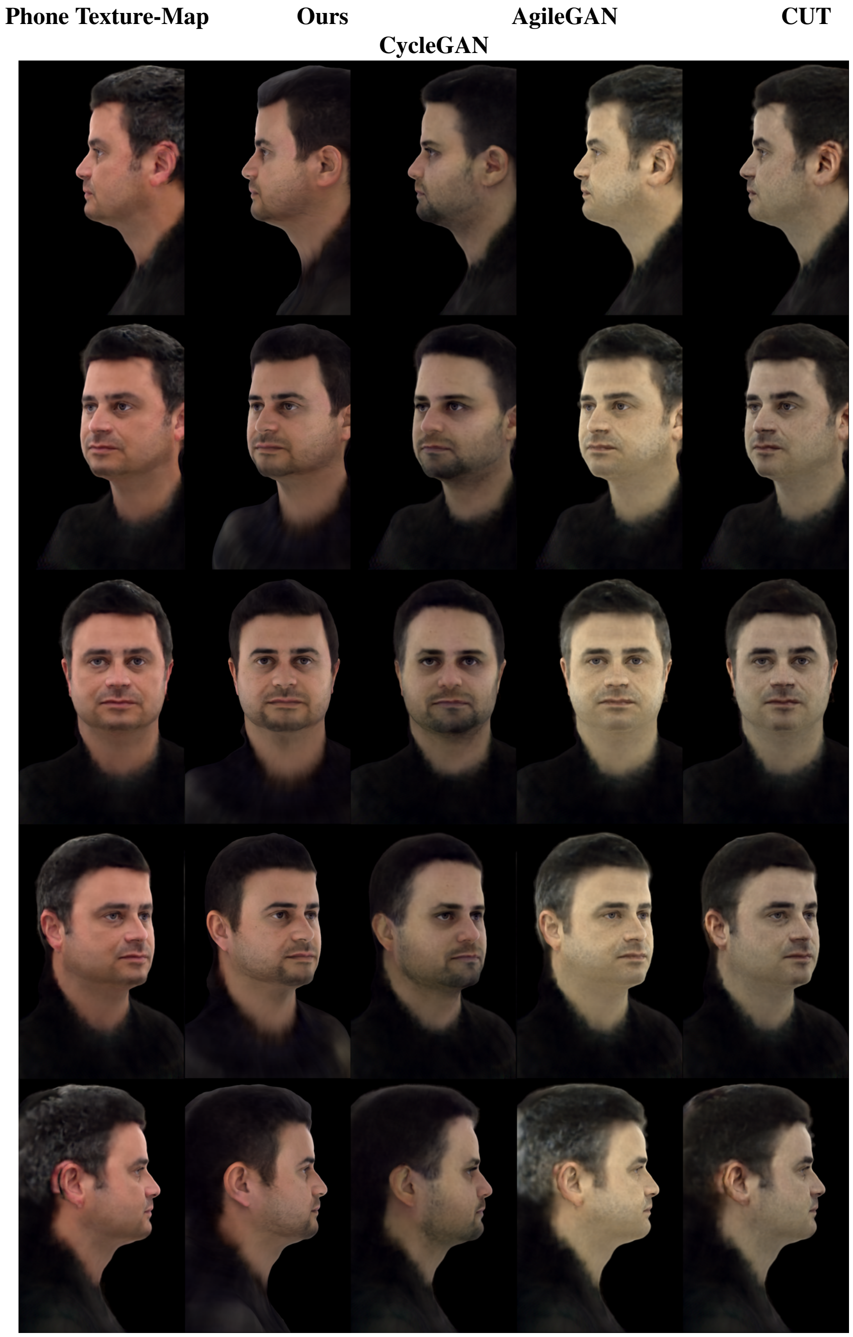

Fig. 4: Render using the Universal Prior Model from AVA [1]

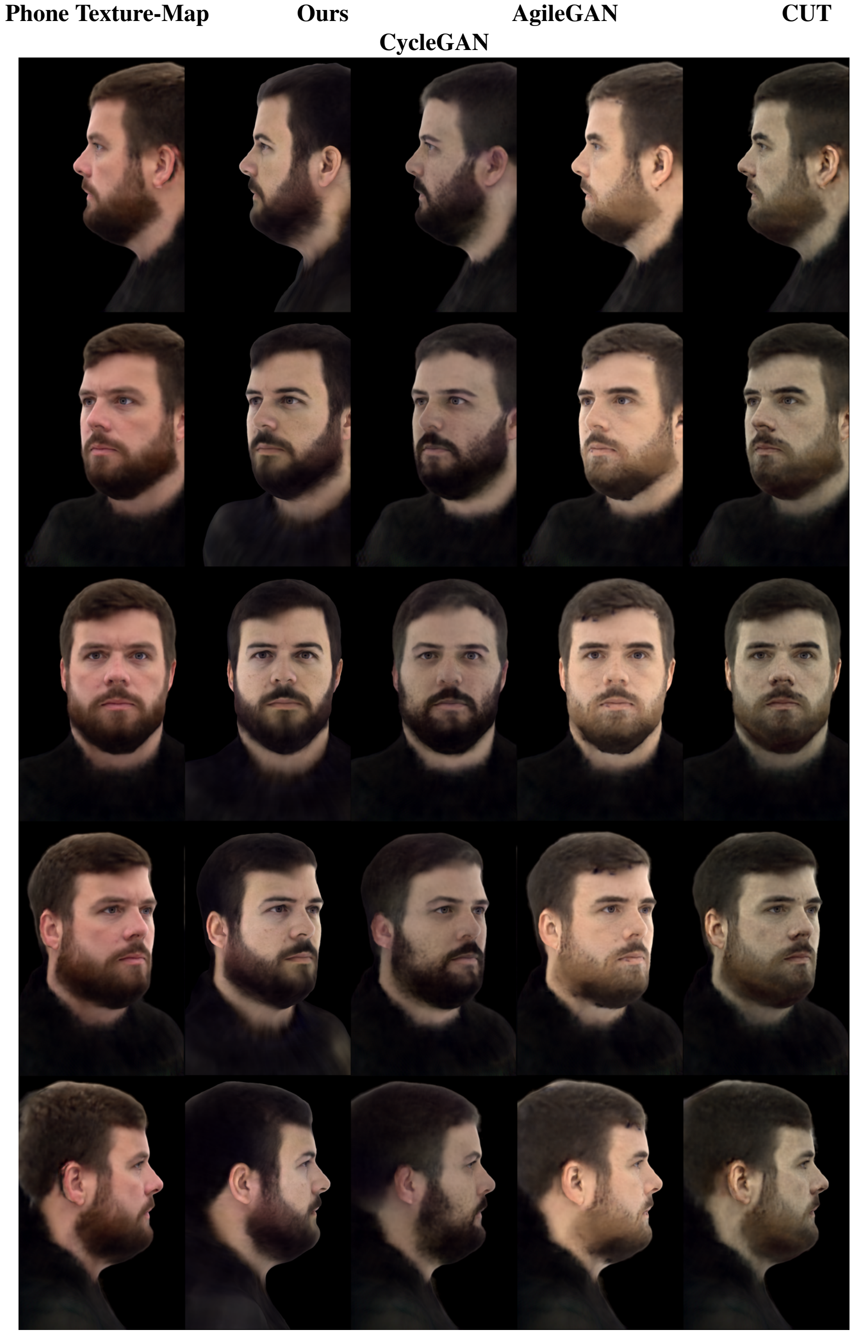

Fig. 5: Render using the Universal Prior Model from AVA [1]

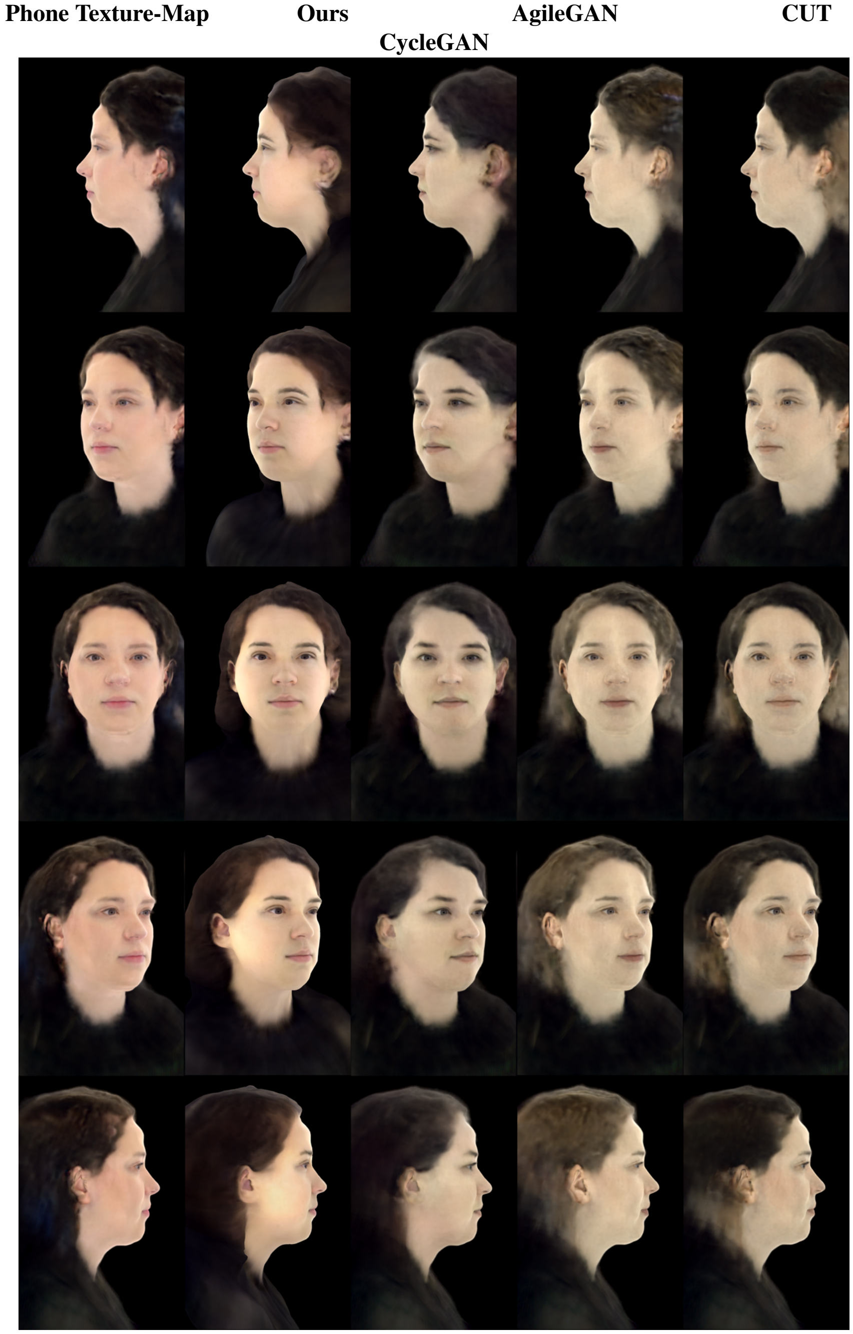

Fig. 6: Render using the Universal Prior Model from AVA [1]

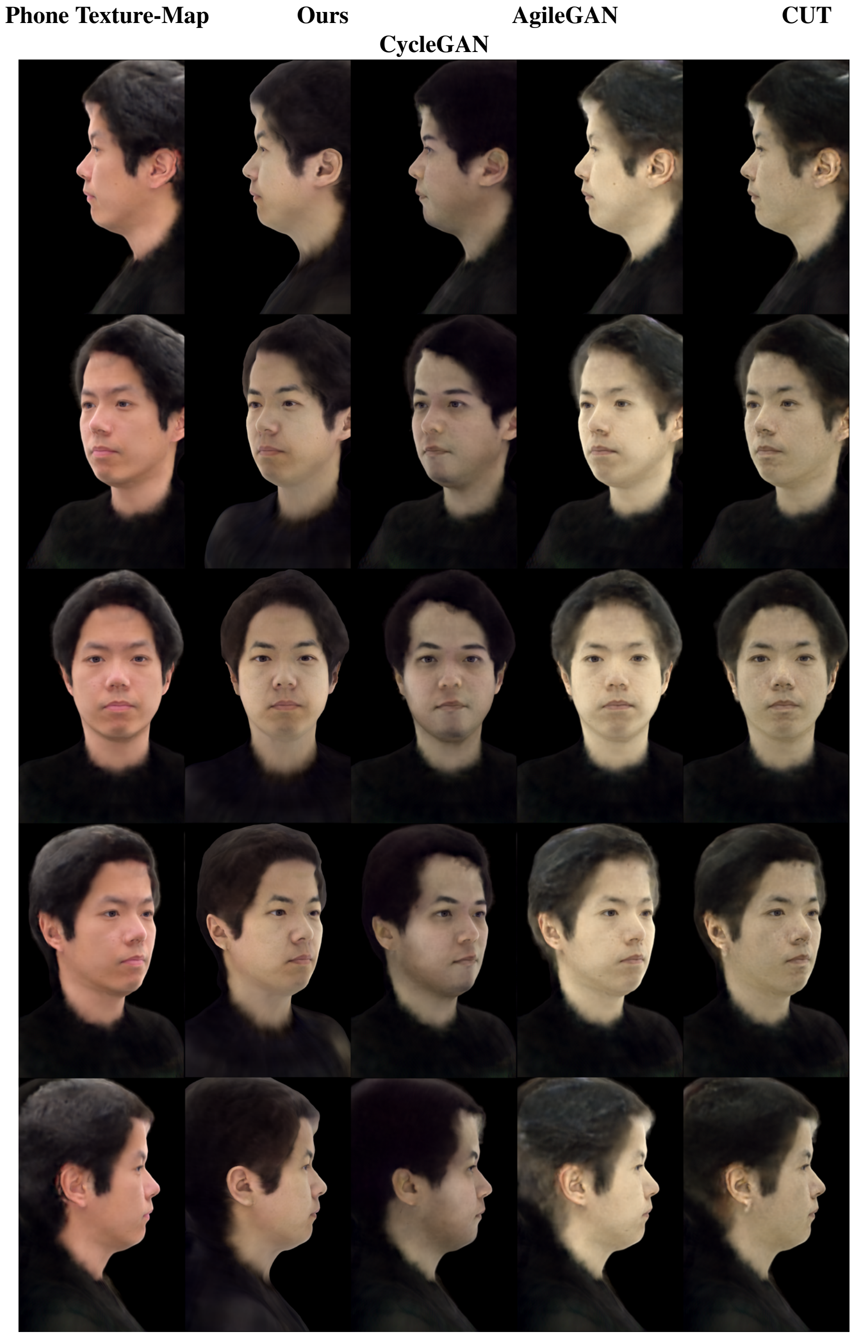

Fig. 7: Render using the Universal Prior Model from AVA [1]

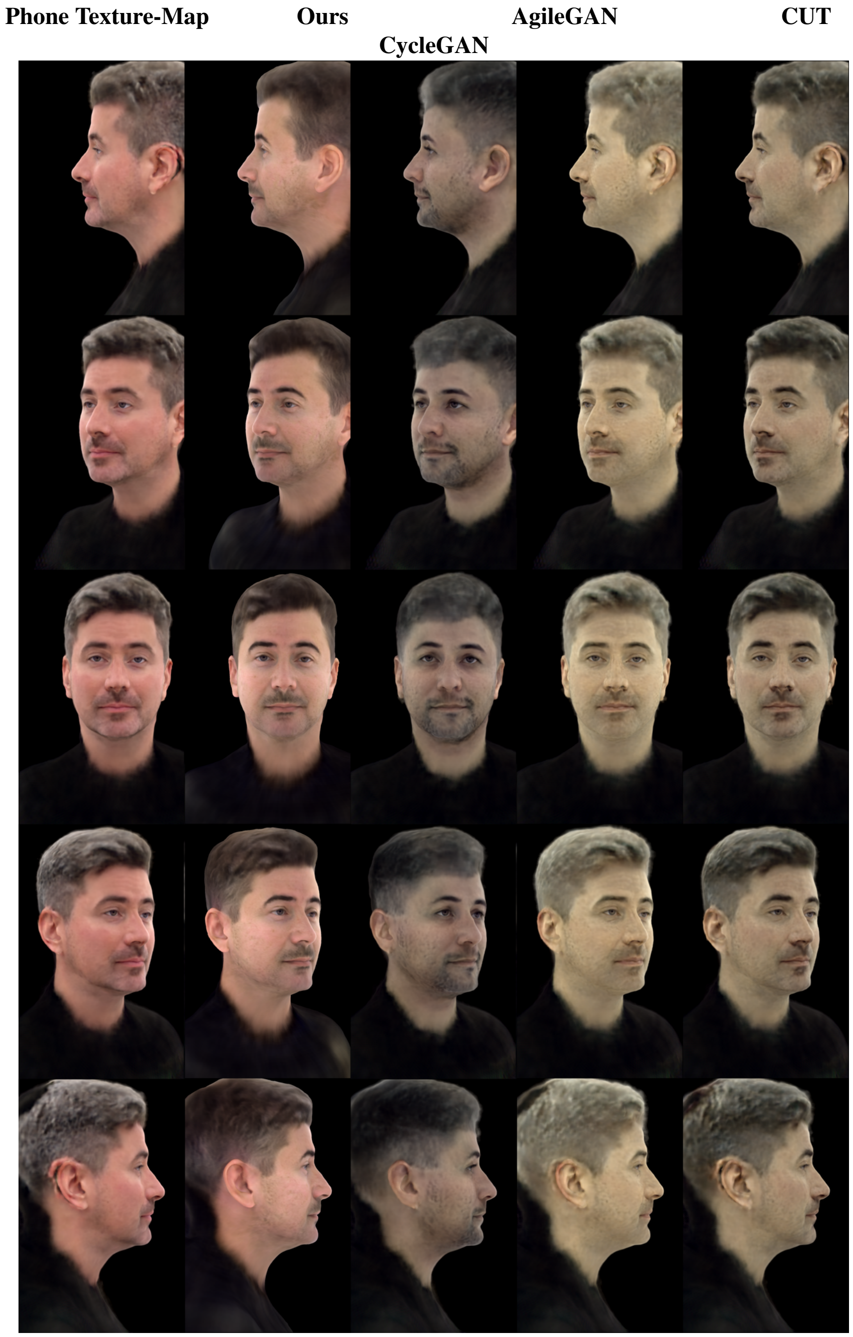

Fig. 8: Render using the Universal Prior Model from AVA [1]

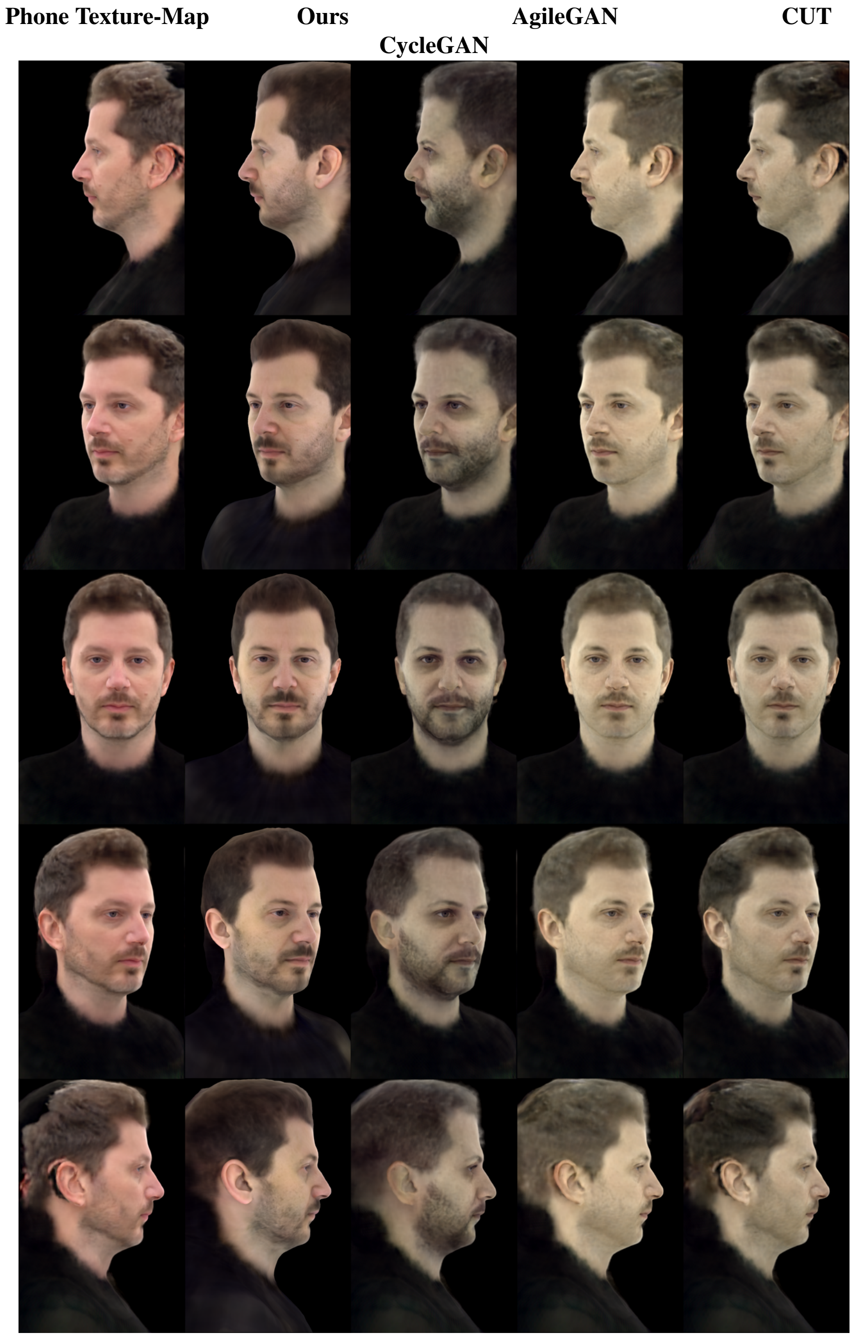

Fig. 9: Render using the Universal Prior Model from AVA [1]

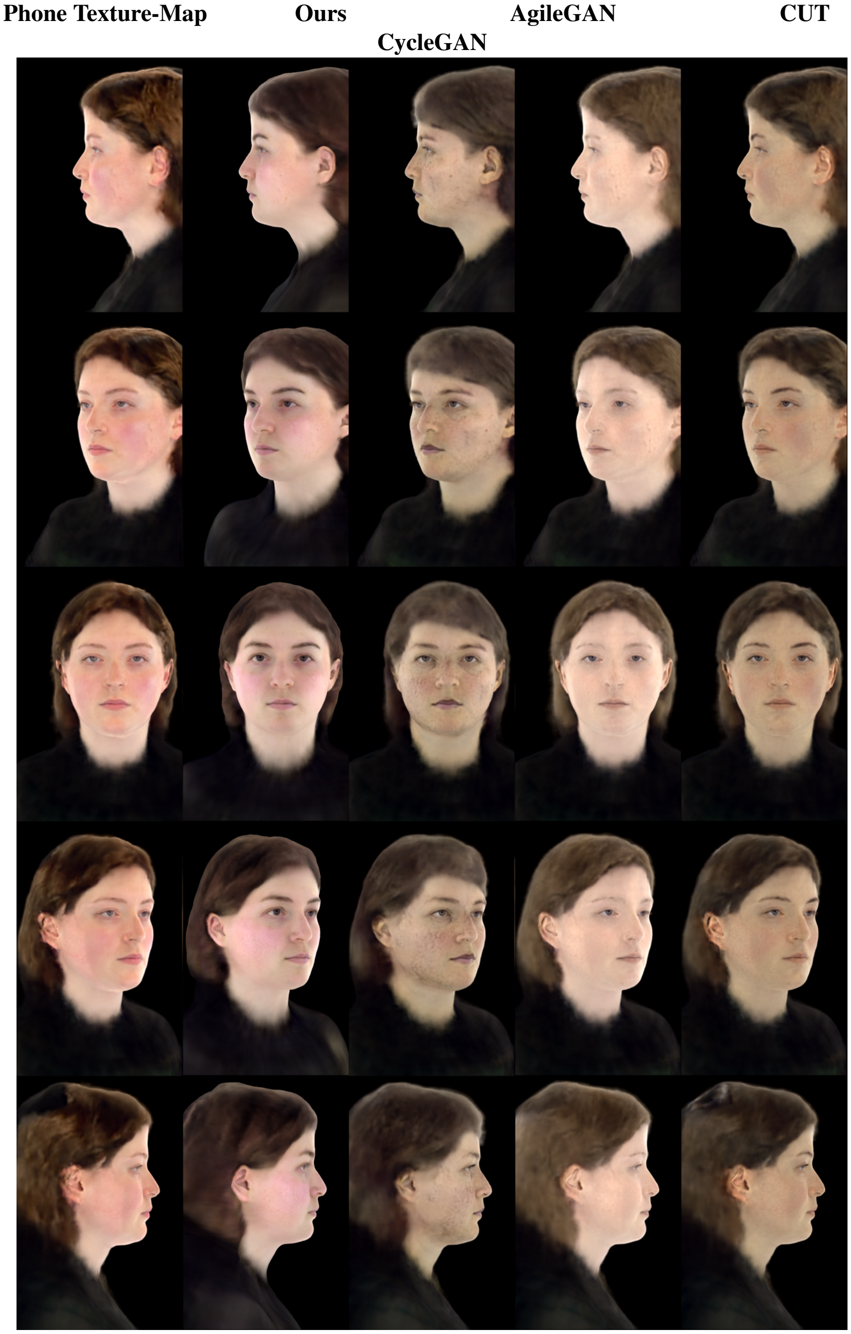

Fig. 10: Render using the Universal Prior Model from AVA [1]

preservation. In Table 1, we ablate this intuition. We compare optimizing the full network (‘Full Network Opt’), optimizing after the 8×8 resolution and optimizing after the 16×16 resolution (‘Opt16’). To measure identity preservation we use the FaceID metric and to measure the fidelity to the target distribution we use the Kernel Inception Distance (KID) [2] metric. As can be seen, optimizing after the 8×8 resolution gives us the best balance between identity preservation and fidelity to the target distribution. Optimizing the whole network (‘Full Network Opt’) yields a better KID but at the cost of a significantly worse identity preservation (a higher FaceID). On the flip side, optimizing after the 16×16 (‘Opt16’) results is a marginal improvement in FaceID but significantly worse KID.

3 Ablation of $\mathcal{L}_{FaceID}$ and $\mathcal{L}_{Percp}$

In this section, we ablate $\mathcal{L}_{FaceID}$ and $\mathcal{L}_{Percp}$ from Eq. (1). As can be seen in Fig 11 and Table 2, not using $\mathcal{L}_{FaceID}$ and $\mathcal{L}_{Percp}$ severely degrades performance and leads to non-convergent results. Using $\mathcal{L}_{Percp}$, leads to significant improvements in training stability and results but there is still an identity shift. Using both $\mathcal{L}_{FaceID}$ and $\mathcal{L}_{Percp}$ yields the best results with the identity strongly preserved after the transfer to studio-like lighting and inpainting of the missing regions.

Models	Full Loss	w/o $\mathcal{L}_{FaceID}$	w/o $\mathcal{L}_{FaceID}$ and $\mathcal{L}_{Percp}$
FaceID ↓	$5.36e-4$	$1.33e-3$	$2.79e-3$

Table 2: Ablation of Losses. Best and Second Best scores are highlighted. Note: These are the results of $\mathcal{G}_{Studio}$ without the facial details added by our diffusion model

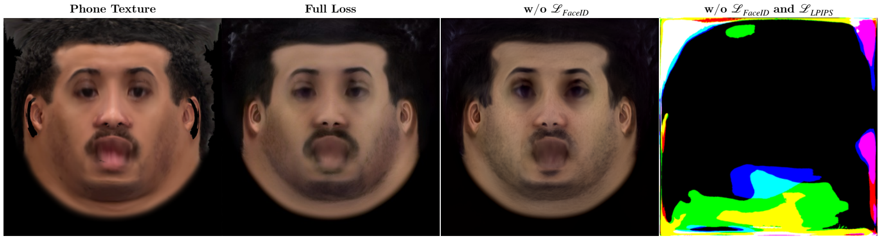
Phone Texture
Full Loss
w/o $\mathscr{L}_{FaceID}$
w/o $\mathscr{L}_{FaceID}$ and $\mathscr{L}_{LPIPS}$

Fig. 11: Ablation of FaceID and LPIPS loss: As can be seen, not using $\mathcal{L}_{FaceID}$ and $\mathcal{L}_{Percp}$ leads to a serve deterioration in performance. Not using $\mathcal{L}_{FaceID}$ causes a significant identity shift. Using the full loss leads to the best results. Note: These are the results of $\mathcal{G}_{Studio}$ without the facial details added by our diffusion model.

4 Ablation of $\mathcal{L}_{Percp-Recons}$

In this section, we ablate $\mathcal{L}_{Percp-Recons}$ from Eq. (1). While $\mathcal{L}_{FaceID}$ and $\mathcal{L}_{Percp}$ help preserve identity and lead to more stable training, they're do not penalize global skin-tone shifts that may occur when an in-the-wild texture map is transferred to studio-like lighting as can be seen in Fig 12. Using $\mathcal{L}_{Percp-Recons}$ with a very small amount of data helps prevent this as it forces $\mathcal{G}_{Studio}$ to maintain skin-tone while transferring from in-the-wild lighting to studio-like lighting.

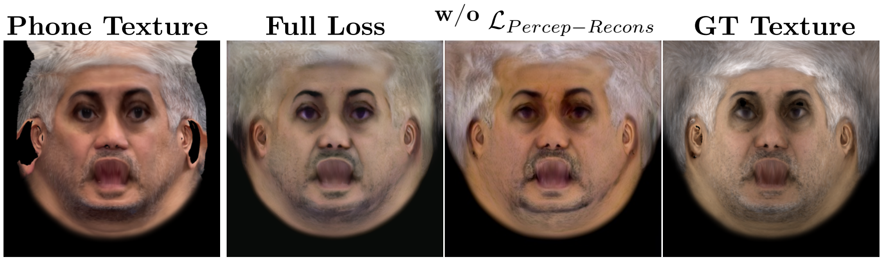

Fig. 12: Ablation of $\mathcal{L}_{Percp-Recons}$: Not using $\mathcal{L}_{Percp-Recons}$ leads to a minor shift in skin-tone.

5 Limitations

As mentioned in the paper, our method only models the head, important regions such as the shoulder and torso and not modelled and are left for future work. Additionally, as shown in Fig 13, our method cannot handle head accessories well, this is because the studio-captured data does not have any head-accessories.

References

1. Cao, C., Simon, T., Kim, J.K., Schwartz, G., Zollhoefer, M., Saito, S.S., Lombardi, S., Wei, S.E., Belko, D., Yu, S.I., Sheikh, Y., Saragih, J.: Authentic volumetric avatars from a phone scan. ACM Trans. Graph. (2022)
2. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: Proc. NeurIPS (2020)

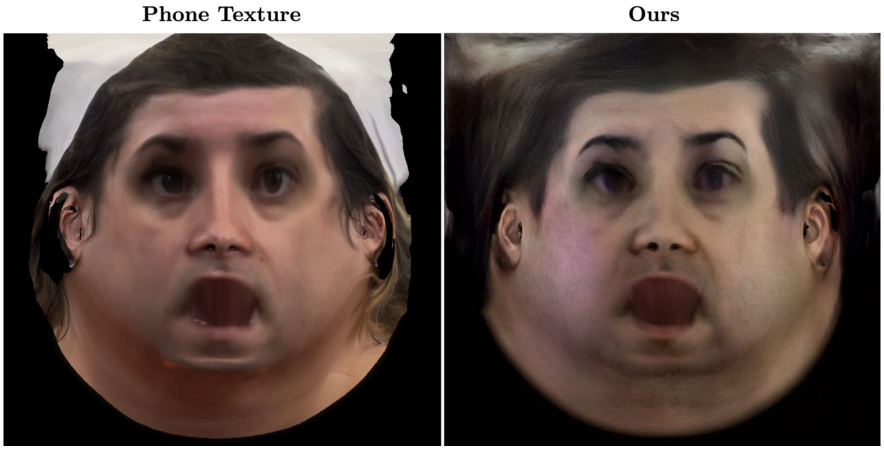

Fig. 13: Limitations: Our method fails to preserve head-accessories