Instant Volumetric Head Avatars
–Supplemental Document–

Wojciech Zielonka Timo Bolkart Justus Thies
Max Planck Institute for Intelligent Systems, Tübingen, Germany
{wojciech.zielonka, timo.bolkart, justus.thies}@tuebingen.mpg.de

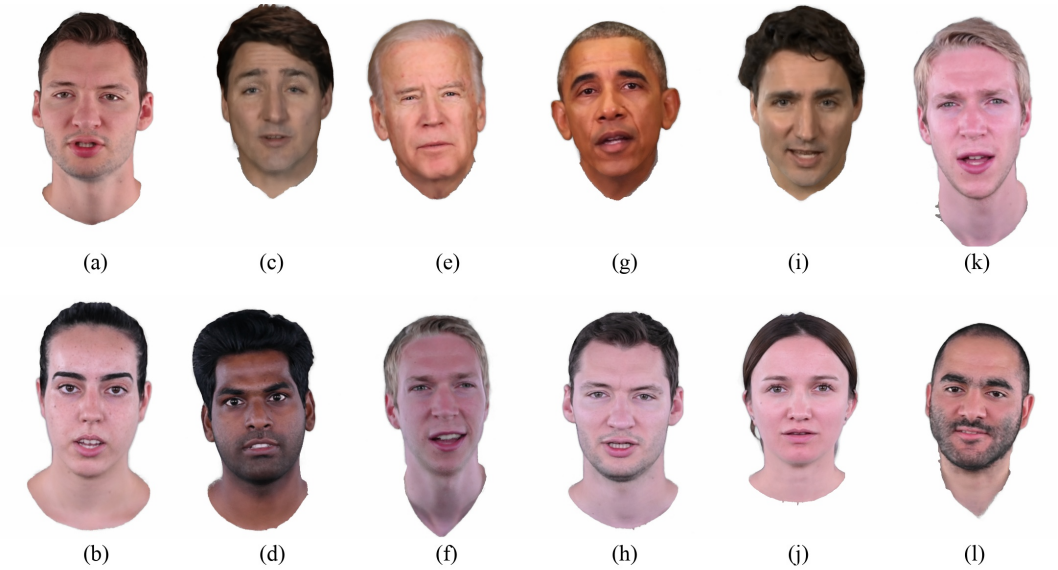

Figure 1. Our dataset consists of twelve sequences obtained from both in-house recordings and YouTube. Here, we present their respective volumetric avatars which INSTA optimizes in less than 10 minutes. As can be seen, our method works well in both cases producing photorealistic videos even for in-the-wild sources (c, e, g, i). Please see the supplemental video for animations of these avatars.

This document elaborates on additional results and potential applications beyond volumetric video conferencing. Specifically, we cover details about the acceleration structures for ray sampling with respect to the neural graphic primitives [5] (see Section 1). We demonstrate facial expression transfer in Section 2. In Section 3, we show additional qualitative and quantitative results in terms of predicted normal and error maps.

1. Implementation Details

Accelerated Ray Marching. INSTA is based on NGP [5], which achieves significant speedup by adapting the sampling strategy utilizing occupancy grids. For a given scene, a separate gird of 128^3 is used to store an occupancy bit. During ray marching for a given sample point, a bit value is measured to determine if this position should be skipped. In this way, samples in empty spaces can be effectively omitted. The occupancy grid is continually updated during the training based on the density values that can be predicted from the neural graphic primitives. To accommodate dynamic scenes, we adapted this mechanism. Specifically, we construct the acceleration structure in the deformed space where we shoot rays, which is different from the canonical space where the neural graphic primitives are learned. Throughout the training, the acceleration structure converges to a Boolean union across all expressions in the training dataset. Optionally, during the update step of the acceleration structure, the occupancy bit for voxels that lay on the isosurface around the canonical mesh can be manually set to on thus creating a fixed sampling region. This approach impedes the rendering time. However, sequences with very expressive motion can benefit from it, especially for motion extrapolation. In Table 1, we detail the hyperparameters of the hashing grid used to store the radiance field.

FLAME masks. Our method uses a mask defined on the simplified FLAME topology complemented with an additional set of triangles for the mouth cavity which is used for

Parameter	Value
Number of levels	16
Hash table size	2^{18}
Number of features per entry	8
Coarsest resolution	16
Finest resolution	2048

Table 1. Hyperparameters used for the hash encoding grid [5].

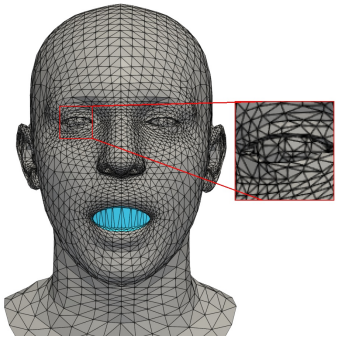

Figure 2. For the expression conditioning, we only consider the mouth region as depicted in magenta color. Moreover, we simplify the eyeballs, since they are modeled as densely tessellated spheres which introduce unnecessary computation overload for nearest neighbor search.

expression conditioning (Section 3, main paper). Figure 2 depicts the conditioning region which handles the dynamic appearance of the mouth interior. Additionally, the new eyeballs topology is magnified. For the mesh simplification, we used Garland et al. [2] to compute a new set of faces which we later transferred to all meshed for a given sequence.

2. Applications

Our volumetric avatars have many potential applications. Because they are controlled by a parametric face model, the expression transfer can be easily applied. We follow NeRFace [1] and calculate relative expressions by manually selecting a neutral face of the target $T_{neutral}$ and source $S_{neutral}$ for which we calculate delta expressions $\Delta_i = S_i - S_{neutral}$. Finally, those relative delta vectors for each frame can be transferred to the target by $T_i = T_{neutral} + \Delta_i$. This step is necessary due to the fact that 3DMMs do not completely disentangle identity and expressions, thus transferring directly the expression coefficients from one actor to the other will change the mesh shape.

Expression transfer has been demonstrated in a variety of state-of-the-art methods on facial avatar reconstruction [1,3, 7, 8, 10]. Specifically, the expression from a source subject is captured and then applied to a facial avatar of a different person. Figure 3 shows that our method can be used for such an application.

(a) (b) (c) (d) (e)

Figure 3. Expression transfer: (a) source actor, (b-e) target subjects with expression from (a).

3. Additional Results

In Figure 6, we present estimated geometry as normal maps. NeRF [4] can recover geometry, but it contains noise. Therefore, NeRFace [1], which is based on NeRF, inherits the same problem. In contrast, NHA [3] uses an explicit mesh representation based on the FLAME topology. However, NHA produces deformed geometry, especially, for the ears. Moreover, the optimized geometry is low-quality and misses many details, which are compensated by neural textures in the final image synthesis. IMAvatar [10] is able to recover high-quality geometry. The approach, based on IDR [9], can take up to several days to optimize the dynamic occupancy field. In contrast, our approach only needs a fraction of this time to optimize an avatar. In the face region, the geometry quality is on par with IMAvatar. How-

Method	Time	Units	GPUs
IMAvatar [10]	~ 4	day	1
NeRFace [1]	~ 3	day	1
NHA [3]	~ 13	hour	3
Ours	~ 10	**minute**	1

Table 2. Average training times for the avatar creation at a resolution of 512^2, except IMAvatar which is using 256^2. Note that the dataset generation for each of the methods is not taken into a consideration and only the avatar training time is measured.

ever, at the hair region where we do not have access to a geometric prior, the quality is similar to NeRFace. In addition to the normal maps, we show the photo-metric error for a single frame in Figure 4 using an RGB-base ℓ_1 metric; a perceptual evaluation for the entire sequences is shown in Figure 5.

4. Training Time Evaluation

In Table 2, we present the average training times for each method. Our method uses a local gaming PC equipped with a modern GPU Nvidia RTX 3090 and requires about 10 min to reconstruct a volumetric avatar with high-frequency details. For the baselines, we use their original configurations on a compute cluster. Specifically, an Nvidia Quadro 6000 was used for the single GPU methods [1, 10], and for NHA [3], three Nvidia A100 40GB GPUs were used. While running on commodity hardware, our method is orders of magnitude faster than the others, making it more versatile and energy-saving.

5. Broader Impact

INSTA synthesizes photo-realistic volumetric avatars from monocular RGB images and can extrapolate to novel views utilizing the 3DMM geometry prior. Since INSTA does not require sophisticated capture setups, it can be applied to standard videos that can be captured with a webcam or a smartphone or downloaded from YouTube. While our research focuses mainly on connecting people via teleconferencing, there is a risk of misuse. Specifically, our method could be abused to produce so-called DeepFakes, which can be used for misinformation, cyber mobbing, identity theft, or other harmful criminal acts. Unfortunately, we are not able to prevent the misuse of our technology. However, conducting research openly and transparently could raise awareness of nefarious uses. We will share our codebase to enable research on digital multi-media forensics, where synthesis methods are needed to produce a training corpus for forgery detection [6].

All participants of our in-house recordings in the study have given written consent to the usage of their video material for this publication. YouTube videos were taken from public domains.

References

[1] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8649–8658. Computer Vision Foundation / IEEE, 2021. 2, 3

[2] Michael Garland and Paul S. Heckbert. Surface simplification using quadric error metrics. pages 209–216, 1997. 2

[3] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular RGB videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18632–18643. IEEE, 2022. 2, 3, 4

[4] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 405–421. Springer, 2020. 2, 6

[5] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 1, 2

[6] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. *ICCV 2019*, 2019. 3

[7] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6):183:1–183:14, 2015. 2

[8] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of RGB videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2387–2395. IEEE Computer Society, 2016. 2

[9] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2, 6

[10] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I M avatar: Implicit morphable head avatars from videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 13535–13545. IEEE, 2022. 2, 3, 4, 6

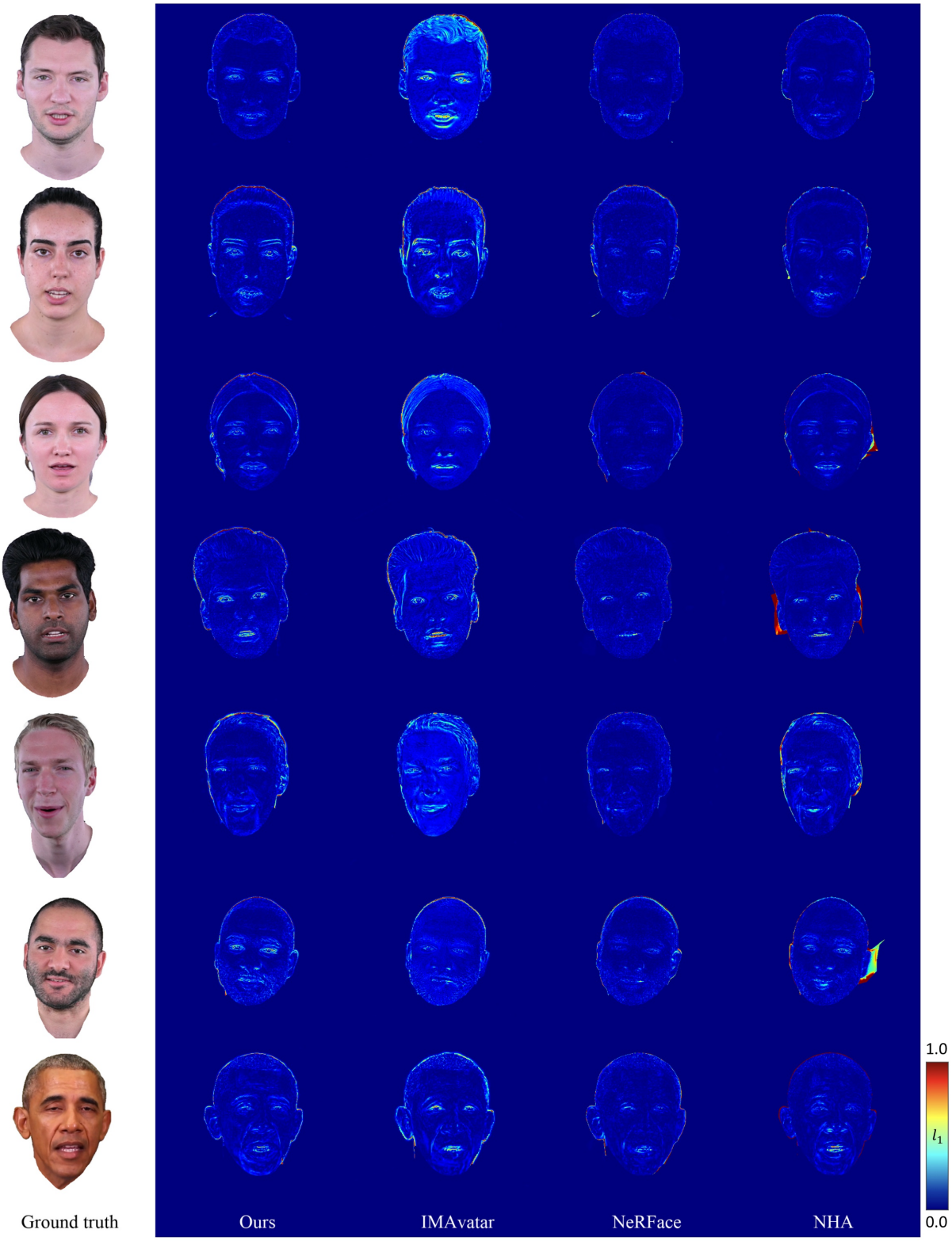

Figure 4. The heatmaps based on l_1 RGB distance represent photo metric errors on the test sequences. IMAvatar [10] synthesizes images with a low level of detail. Geometry mispredictions of NHA [3] create artifacts around the ear region.

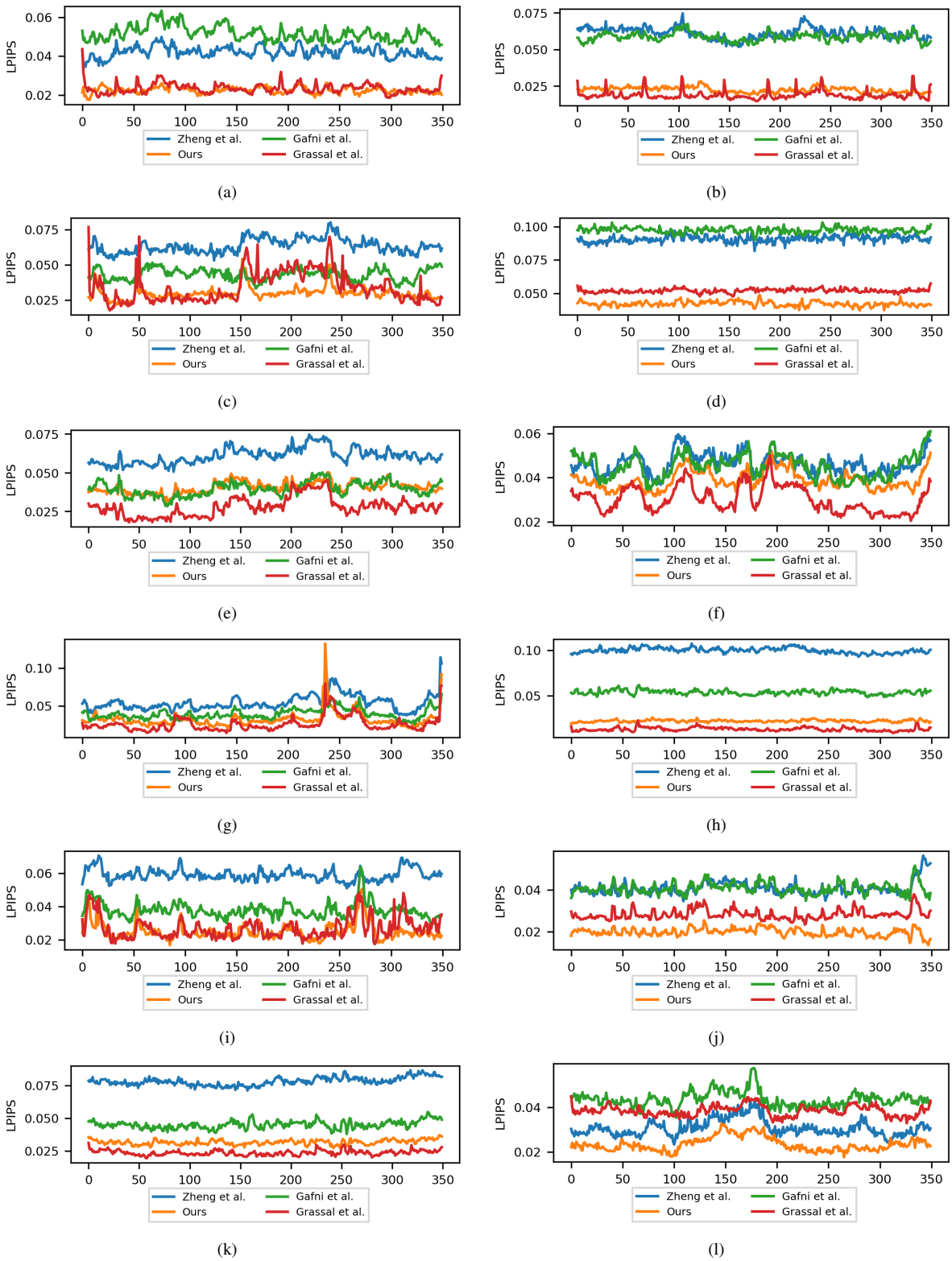

Figure 5. Evaluation of the perceptual error for each of the volumetric avatars from our dataset on the test sequence. The alphanumerical order matches Figure 1. Our method achieves the lowest errors for color reconstruction and captures well even high-frequency details like freckles or wrinkles.

Figure 6. NeRF [4] produces noisy normals maps, which can be seen in the results of Gafni et al. Our method uses additional geometric prior, which helps reduce the noise in specific areas, however, regions like hair are still problematic. The best results are achieved by IMAvatar [10], which is based on a modified version of the IDR [9] approach. However, it takes a few days to achieve this result, while ours reaches similar quality, especially, in the face region, in about 10min.