

HERA: Hybrid Explicit Representation for Ultra-Realistic Head Avatars

Supplementary Material

In the supplementary material, we provide more validations about HERA, including comparisons, parameter quantity and inference efficiency. Then, we introduce two additional tasks to validate our hybrid representation further. For more visual demonstrations, please refer to our video.

A. More Validations

As detailed in the Experiments section of the main paper, we implement a baseline that relies exclusively on texture-mapped meshes, called mesh based avatars. We compare our HERA with the baseline and GaussianAvatars [13] across various perspectives, expressions, and poses, as illustrated in Fig. 1. HERA integrates the advantages of both mesh and 3DGS representations while also overcoming their limitations. HERA takes advantage of the accurate rendering of human faces achieved through UV-mapped mesh, while also utilizing the modeling features for hair and eyelashes available in 3DGS. In contrast, mesh representation struggles to reconstruct complex structural shapes, while 3DGS is less adept at modeling low-frequency geometric surfaces with high-frequency textures.

Furthermore, our HERA utilizes an average of 129,697 splats to model an avatar from Multiface dataset, which is half the number used by GaussianAvatars (an average of 261,783 splats). Even if considering the mesh parameters, HERA uses fewer parameters to create a more realistic avatar, which states the proposed hybrid representation makes different primitives reconstruct the scene more efficiently. To infer a video at 2048×1334 resolution on a single NVIDIA RTX A100 GPU, HERA renders at 81 FPS which suffices for real-time applications.

B. Novel View Synthesis on Static Scenes

Overview. To further validate the effectiveness of our proposed hybrid presentation, we conduct an experiment on novel view synthesis in static scenes. For this, we follow the settings outlined in 3DGS [8]. It is important to note that we do not rig the 3D Gaussians onto the mesh facet in this scenario, which means that the association between the two types of primitives is solely established through the hybrid rendering pipeline during the optimization process.

Datasets. For evaluating static scene data, we perform experiments on the Tanks and Temples [9] and Mip-NeRF 360 [2] datasets. We utilize 5 scenes from the Tanks and Temples dataset, along with all scenes from the Mip-NeRF 360 dataset in our experiments.



Figure 1. Comparisons of free views, expressions and poses on Multiface dataset [15]. From left to right, we display the results of mesh based avatars, GaussianAvatars [13] and ours, respectively. Since the viewpoints are freely selected, there are no ground truth images for reference. Zoom in for better views.

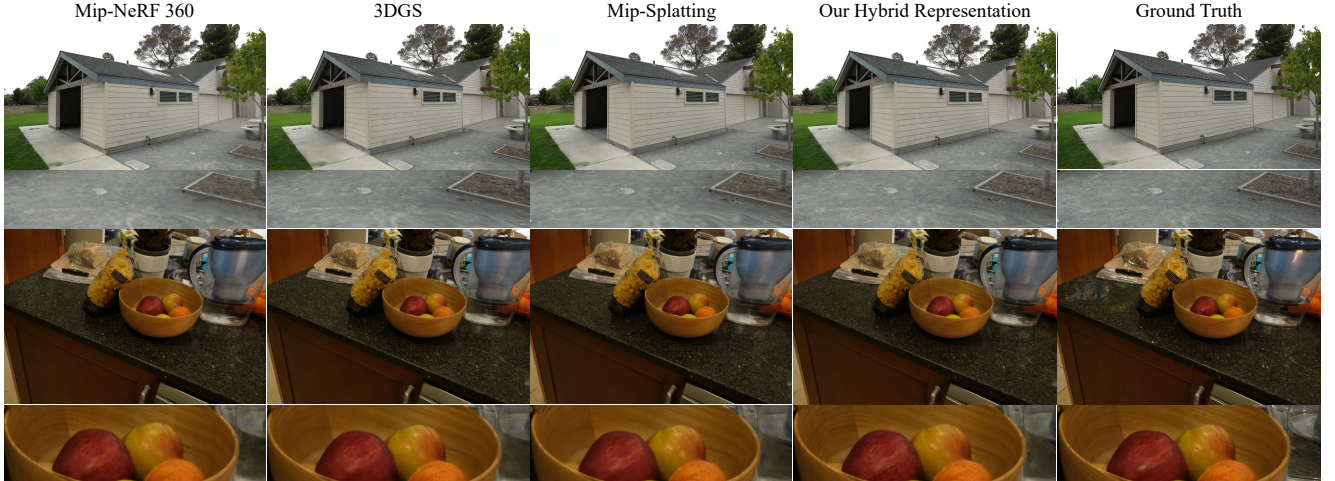


Figure 2. Novel view synthesis on Tank and Temples [9] (top row) and Mip-NeRF 360 [2] (bottom row) datasets. From left to right, we display the results of Mip-NeRF 360 [2], 3DGS [8], Mip-Splatting [17], our hybrid representation and ground truth images, respectively. Zoom in for better views.

Method	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
NeRF [11]	23.85	0.605	0.451
Mip-NeRF [1]	24.04	0.616	0.441
Plenoxels [6]	23.08	0.626	0.463
Instant-NGP [12]	25.68	0.705	0.302
Mip-NeRF 360 [2]	27.57	0.793	0.234
Zip-NeRF [3]	28.54	0.828	0.189
3DGS [8]	27.70	0.826	0.202
Mip-Splatting [17]	27.79	0.827	0.205
Our Hybrid Representation	27.61	0.828	0.199
Our Hybrid Representation*	27.75	0.829	0.194

Table 1. The quantitative results on Mip-NeRF 360 dataset [2]. The * indicates incorporating our hybrid representation with Mip-Splatting [17]. We denote the best, second best, and third best scores in different colors.

Baselines. We select Mip-NeRF 360 [2], 3DGS [8] and Mip-Splatting [17] as the primary comparative baselines. Additionally, NeRF [11], Mip-NeRF [1], Plenoxels [6], Instant-NGP [12] and Zip-NeRF [3] are used for quantitative comparisons.

Comparisons. The quantitative results for Mip-NeRF 360 dataset are presented in Table 1. Our hybrid representation demonstrates significant improvements in LPIPS metrics, achieving comparable performance in PSNR and SSIM metrics relative to state-of-the-art methods. The quantitative results for Tanks and Temples dataset are provided in Table 2. Our representation outperforms the others, showing marked improvements in LPIPS metrics. This can be attributed to the more intricate geometries found in the Mip-NeRF 360 dataset compared to the Tanks and Temples

Method	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Plenoxels [6]	21.94	0.708	0.386
Instant-NGP [12]	22.26	0.724	0.347
Mip-NeRF 360 [2]	23.61	0.765	0.283
3DGS [8]	24.01	0.814	0.228
Mip-Splatting [17]	24.21	0.822	0.216
Our Hybrid Representation	24.29	0.822	0.207
Our Hybrid Representation*	24.39	0.824	0.206

Table 2. The quantitative results on Tanks and Temples dataset [9]. The * indicates incorporating our hybrid representation with Mip-Splatting [17].

dataset. Many scenes in Mip-NeRF 360 dataset feature extensive grassland areas at the image’s center, while the target objects in Tanks and Temples dataset possess smoother surfaces, making them more suitable for mesh modeling.

The qualitative results for static scenes are illustrated in Fig. 2. Our approach shows significant enhancements on smooth surfaces characterized by complex color variations. However, Mip-Splatting still exhibits rendering defects on smooth surfaces, suggesting that these issues are not solely related to anti-aliasing. These experimental results underscore the effectiveness of our proposed hybrid representation. Further qualitative results from the Mip-NeRF 360 and Tanks and Temples datasets are displayed in Fig. 3, demonstrating that improvements are most pronounced on smooth surfaces, which are effectively represented using meshes.

C. Dynamic Scene Reconstruction

Our hybrid representation can also model Freeview videos of dynamic scenes. Given a tracked non-parametric mesh,



Figure 3. More qualitative results on the static scene datasets [2, 9]. From left to right, we display the results of 3DGS [8], Mip-Splatting [17], our hybrid representation and ground truth images, respectively.

we design a model that utilizes the proposed hybrid representation to jointly optimize the parameters of mesh and 3DGS. As the non-parametric mesh always be incomplete to represent the holistic motion, we apply the tri-plane feature grid to model the deformation field by following 4D-GS [14], instead of rigging splats to the mesh as done in HERA.

C.1. Methods

Given the 3D coordinates of a Gaussian, we represent its position, scale, rotation, and opacity in the temporal space by utilizing the deep neural network and tri-plane feature grid [4, 5, 7]. Specifically, given a Gaussian splat in the canonical space, we apply a multi-resolution HexPlane [4] and shallow MLP ϕ to extract and decode the features from a 4D K-Planes module [7] which contains 6 feature planes according to the position in canonical space $\mathbf{x}_c = (x, y, z)$ and time t . We decode the position, scale, rotation, and opacity of the 3D Gaussian at any given point in the temporal sequence, thereby enabling a robust and detailed representation of the dynamic scene deformation field modeling.

Specifically, the extracted features on the spatial and temporal space are denoted as

$$\begin{aligned} D(x, y, z, t) = \{ & D_l(x, y), D_l(y, z), D_l(x, z), \\ & D_l(x, t), D_l(y, t), D_l(z, t) | l \in \{1, 2\} \}, \end{aligned} \quad (1)$$

where l is the upsampling scale and $D_l(i, j)$ is the multi-resolution feature plane.

After the features $D_l(i, j)$ from multiple feature planes are obtained, they are fused to a global feature vector which consists of both temporal and spatial information:

$$\begin{aligned} \mathcal{H} = \bigcup_i \prod \text{interp}(D_l(y, z), D_l(x, z), \\ D_l(x, t), D_l(y, t), D_l(z, t)), \end{aligned} \quad (2)$$

where the "interp" indicates the bilinear interpolation for the 2D feature grid plane. Take the position parameter \mathbf{x} of 3D Gaussian as an example, the deformed position can be computed by decoding the feature vector: $\mathbf{x}(t) = \mathbf{x}_c + \phi_{\mathbf{x}}(\mathcal{H})$. $\phi_{\mathbf{x}}$ is the shallow MLP for decoding the position deformation of Gaussian splat. This design is also applied to the scale, rotation, and opacity parameters of Gaussian splats.

For the mesh representation, since the deformation of mesh has a different pattern from 3DGS, we design the deformation of mesh as the translation of each vertex instead of applying a grid based field, which is a simpler and more direct approach. In this case, each vertex in the mesh has its displacement vector, which directly translates it from its original position to a new position in the deformed state. The mesh based representation contains the vertices, opacity map, and texture feature map for each frame with a shared topological structure.

In the initial phase of our method, we tackle the challenge of tracking a mesh representation in a dynamic scene using a keyframe based approach. Specifically, we utilize Reality Capture Software to generate a high-quality textured mesh for a selected keyframe. This keyframe mesh serves as the base model against which other frames are registered and aligned. Next, we apply the mesh to a point cloud registration algorithm AMM-NRR [16] to establish the mesh-to-points correspondences and get coarse-tracked meshes to other frames.

Subsequently, we refine the tracked meshes further in the differentiable rendering framework to finetune the tracked mesh of each frame by applying the photometric loss \mathcal{L}_2 ensuring that the rendered image closely matches the actual input images from each frame. And the normal consistency loss \mathcal{L}_{nc} to ensure that the surface normals across the mesh faces align well with those estimated from the input data. During this refinement process, the opacity UV map for all the tracked meshes is set to 1 which indicates full visibility, and the opacity and texture map are not subject to optimization, focusing solely on refining the geometry according to the photometric and normal constraints from the dynamic scene.

The photometric loss for the mesh tracking is defined as the L_2 norm term:

$$\mathcal{L}_2 = \sum_{\mathbf{u} \in \mathcal{U}} \|\mathbf{C}(\mathbf{u}) - \mathbf{C}_{gt}(\mathbf{u})\|_2^2, \quad (3)$$

where the \mathcal{U} is the set of the pixels on the image. The loss function corresponding to this non-rigid tracking phase is:

$$\mathcal{L}_m = \mathcal{L}_2 + \lambda_{nc} \mathcal{L}_{nc}, \quad (4)$$

where the \mathcal{L}_{nc} is the normal consistency loss for smooth regularization to the mesh.

After the non-rigid tracking of meshes, we initialize the point cloud of 3DGS in canonical space according to the tracked mesh of each frame. For the linear layer $y = Wx + b$ of MLP for decoding the deformation feature vector, the learnable parameters W are initialized as normal distribution $\mathcal{N}(0, \epsilon)$, where the ϵ is a small value (1×10^{-4}), and the learnable parameters b are initialized as zero. This initialization of the deformation field ensures that the deformations are close to zero at the beginning of the training phase, allowing the 3DGS initialized for each frame to be positioned close to the surface. The vertices, opacity map, and texture feature map from mesh representation are all optimized in this phase.

The loss function of optimizing the hybrid representation is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_{SSIM} \mathcal{L}_{SSIM} + \lambda_{tv} \mathcal{L}_{tv}, \quad (5)$$

where \mathcal{L}_1 and \mathcal{L}_{SSIM} is the L_1 norm loss and structural similarity loss between the rendered image and ground truth

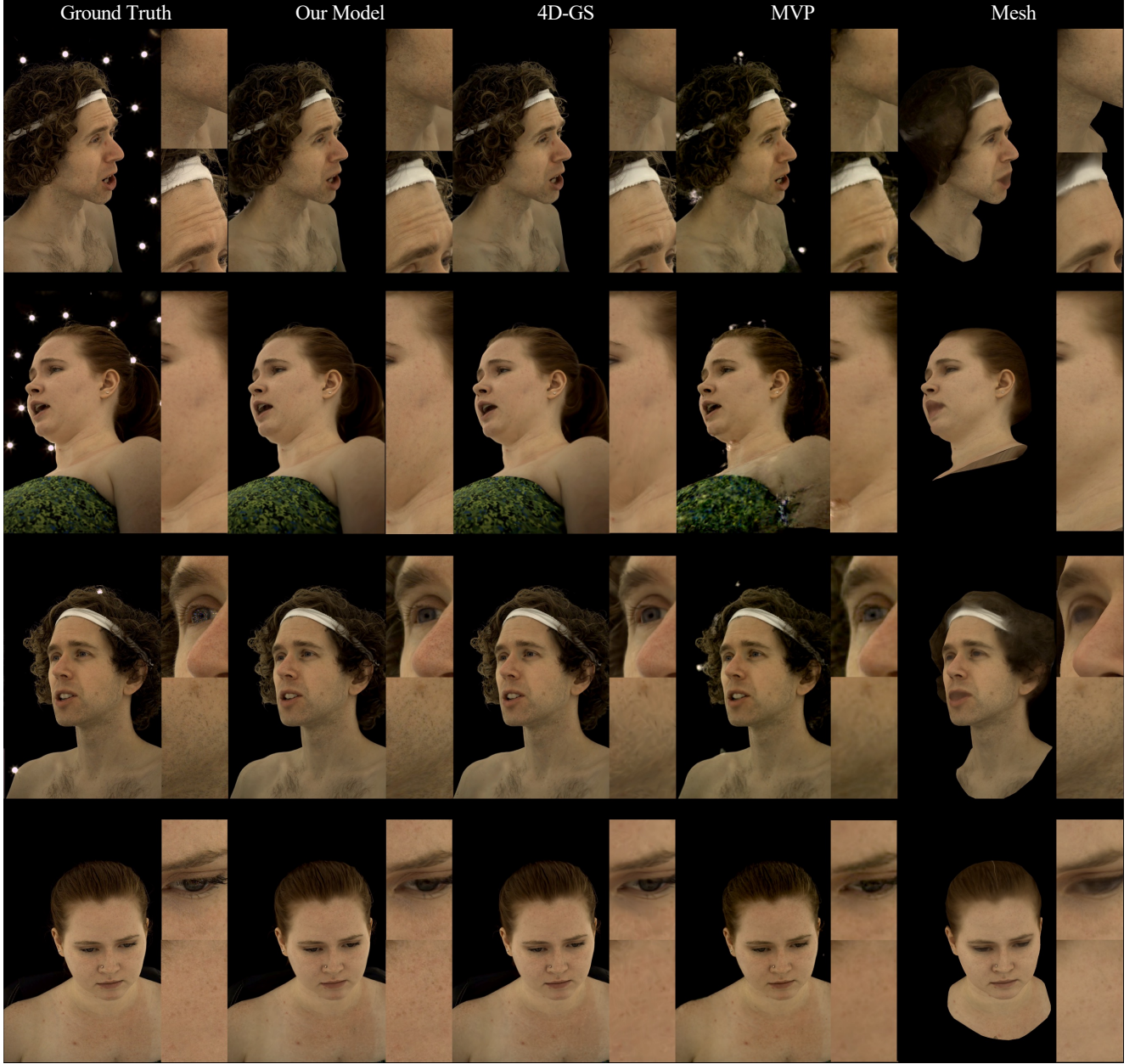


Figure 4. The qualitative experiments on the test views of Multiface dataset [15] on dynamic scene reconstruction task. From left to right, we display ground truth images, the results of our model, 4D-GS [14], MVP [10] and the tracked mesh, respectively. Zoom in for a better view.

image. The \mathcal{L}_{tv} is the Total Variational (TV) loss on the feature plane of the deformation field on the spatial and temporal range, which follows HexPlane. The loss function is applied to the feature plane of the deformation field across both spatial and temporal dimensions to ensure its smoothness.

C.2. Experiments

We train and evaluate our model on the Multiface dataset for the dynamic scene reconstruction task. To validate the robustness and effectiveness of our hybrid representation on this task, we have carried out comparative experiments against 2 baselines: MVP [10] and 4D-GS [14].

The qualitative and quantitative results on the Multiface dataset [15] are shown in Fig. 4 and Table 3, respectively.

Method	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Mesh	21.56	0.6747	0.2759
MVP [10]	32.98	0.8648	0.1612
4D-GS [14]	33.76	0.8789	0.1636
Our Model	34.23	0.8872	0.1320

Table 3. The quantitative comparisons on the Multiface dataset [15] for dynamic scene reconstruction. We denote the **best**, and **second best** scores in different colors.

It can be observed that our method achieves state-of-the-art performance compared with other baselines which indicates its superior performance over existing baselines. This highlights the robustness and effectiveness of our proposed hybrid representation for Freeview video rendering.

Based on modeling the deformation of canonical 3DGS [8], 4D-GS [14] can effectively reconstruct intricate geometries like hair strands with high fidelity. However, despite its strengths in handling complex structures, this method tends to lose subtle details regarding smooth surfaces, such as the fine whiskers on a human face.

Meshes are inherently well-suited for representing smooth surfaces since they allow for modeling the surface over vertices and topology, thereby enabling the capture of minute surface variations.

Our model not only preserves the high-fidelity reconstruction of the hair but also reconstructs the detailed color appearance of the human face. This indicates that our hybrid representation has a significantly better capacity for the details on the smooth surface, which benefits from the mesh representation. The mesh is suitable for representing the smooth surface and the high-resolution texture feature map can achieve excellent modeling for complex color appearance and only occupy very little memory.

References

- [1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5855–5864, 2021. **2**
- [2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5470–5479, 2022. **1, 2, 3**
- [3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19697–19705, 2023. **2**
- [4] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 130–141, 2023. **4**
- [5] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 333–350. Springer, 2022. **4**
- [6] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5501–5510, 2022. **2**
- [7] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12479–12488, 2023. **4**
- [8] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4):139:1–139:14, 2023. **1, 2, 3, 6**
- [9] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (TOG)*, 36(4):78:1–78:13, 2017. **1, 2, 3**
- [10] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhöfer, Yaser Sheikh, and Jason M. Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (TOG)*, 40(4):59:1–59:13, 2021. **5, 6**
- [11] 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. *Communications of the ACM*, 65(1):99–106, 2021. **2**
- [12] 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):102:1–102:15, 2022. **2**
- [13] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20299–20309, 2024. **1**
- [14] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20310–20320, 2024. **4, 5, 6**
- [15] Cheng-hsin Wu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Alexander Hypes, et al. Multiface: A dataset for neural face rendering. *arXiv preprint arXiv:2207.11243*, 2022. **1, 5, 6**

- [16] Yuxin Yao, Bailin Deng, Weiwei Xu, and Juyong Zhang. Fast and robust non-rigid registration using accelerated majorization-minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(8):9681–9698, 2023. [4](#)
- [17] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19447–19456, 2024. [2](#), [3](#)