Topo4D: Topology-Preserving Gaussian Splatting for High-Fidelity 4D Head Capture

Xuanchen Li[1], Yuhao Cheng[1], Xingyu Ren[1], Haozhe Jia[2], Di Xu[2], Wenhan Zhu[3], and Yichao Yan[1†]

[1] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
{lixc6486,chengyuhao,rxy_sjtu,yanyichao}@sjtu.edu.cn
[2] Huawei Cloud Computing Technologies Co., Ltd
{jiahaozhe1,xudi21}@huawei.com
[3] Xueshen AI whzhu@foxmail.com

Abstract. Recent significant advances in high-quality face reconstruction have been made, but challenges remain in 4D face asset reconstruction. 4D head capture aims to generate dynamic topological meshes and corresponding texture maps from videos, which is widely utilized in movies and games for its ability to simulate facial muscle movements and recover dynamic textures in pore-squeezing. The industry often adopts a method involving multi-view stereo and non-rigid alignment. However, this approach is prone to errors and heavily relies on time-consuming manual processing by artists. To simplify this process, we propose Topo4D, a novel framework for automatic geometry and texture generation that optimizes densely aligned 4D heads and 8K texture maps directly from calibrated multi-view time-series images. Specifically, we first represent the time-series faces as a set of dynamic 3D Gaussians with fixed topology in which the Gaussian centers are bound to the mesh vertices. Afterward, we optimize geometry and texture frame-by-frame alternatively for dynamic head capture while maintaining temporal topology stability. Finally, we can extract dynamic facial meshes in regular wiring arrangement and high-fidelity textures with pore-level details from the learned Gaussians. Extensive experiments show that our method achieves superior results than the current SOTA face reconstruction methods in the quality of both meshes and textures. Project page: https://xuanchenli.github.io/Topo4D/.

Keywords: 4D Face Modeling · High Resolution Texture Generation

1 Introduction

4D head capture requires obtaining temporal-continuous topological facial assets, including facial geometries and textures. It has been widely used in entertainment media, such as games, movies, and interactive AR/VR, to create dynamic

†Corresponding author

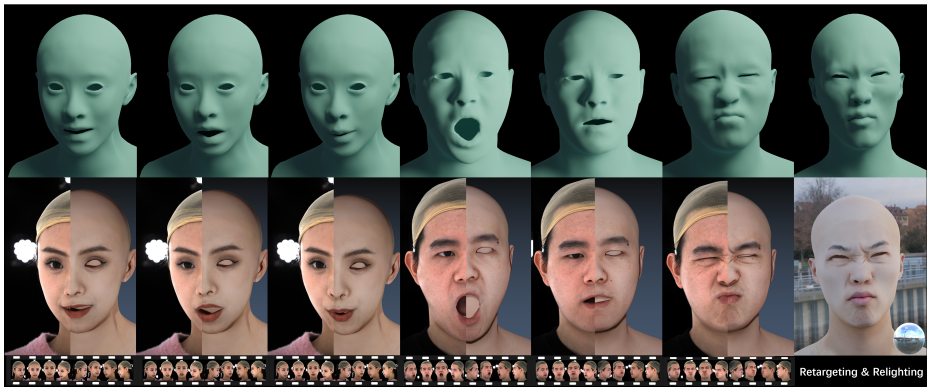

Fig. 1: Example results of our Topo4D . Our method can produce temporal-consistent topological head meshes with high-fidelity 8K textures from calibrated multi-view videos. Captured 4D models can be applied to retargeting and relighting applications.

faces with realistic and immersive quality. The main challenges of 4D head capture are: **1)** representing faces with a fixed **topology** and a regular UV, and **2)** maintaining **temporal stability** among different frames.

To produce captivating and lively 4D facial assets, the industrial pipeline typically employs professional equipment, *e.g.*, Light Stage [19], to capture high-quality multi-view videos. Then multi-view stereo (MVS) [27,47] is used to compute the facial scan of each frame, followed by a non-rigid registration [9] process to superimpose the topologically aligned faces onto the scans. To achieve temporal consistency and obtain usable assets, this process requires marking on the subject's face and a manual post-process by artists. To eliminate the need for manual operations, some methods [8,14,24,65] employ optical flow or other techniques as supervision to automatically warp template models at the expense of processing time. Additionally, they necessitate careful parameter tuning for different subjects to achieve optimal results. Therefore, there is an urgent demand to develop more automated workflows to accelerate the 4D asset reconstruction.

To achieve automatic and efficient facial reconstruction, researchers have developed deep-learning models for **topological** facial asset generation frame-by-frame, which can be divided into two categories. The first line of work [2–4, 7, 21, 26, 28, 36, 52, 57, 59, 60] utilizes parametric models to fit the facial images for mesh creation, where parametric texture or inverse rendering is employed for corresponding texture generation. These methods are highly efficient, but due to their limited expressive abilities, they struggle to produce high-quality textures with diverse identities and complex expressions. Another line of works [3,12,37,43,60,62] proposes to directly regress face models from multi-view images, where a large amount of expensive pre-processed 3D data are utilized for training. Typically, these methods can capture consistent features across multiple views to predict accurate geometry, while several recent works [34,43] successfully generate high-resolution textures with a super-resolution module. Nevertheless,

super-resolution modules often introduce artifacts and fail to faithfully recover the details of faces. Furthermore, these two kinds of aforementioned methods encounter the same predicament in that they are designed to reconstruct each frame individually without directly applying frame continuity, thereby struggling to maintain the temporal coherence between frames.

The recent progress in 3D Gaussian Splatting (3DGS) [31] and its advancements in 4D scene representation [40, 46, 61, 66, 67] bring us inspiration. These methods of high-fidelity 4D scene representation fully take into account the continuity among frames and achieve **temporal-consistent** reconstruction. Moreover, thanks to its advanced rendering pipeline, it can achieve ultra-high resolution texture learning and fast perspective rendering in a memory-efficient manner. Considering their success in 4D scene representation, it would be highly desirable if we could extract high-fidelity facial meshes and textures in the predefined topology from these 3DGS-inspired representations. However, it is nontrivial since the Gaussians in these representations are random and uncontrollable, making it difficult to perfectly register geometries with a fixed topology.

To overcome these challenges, we propose a novel optimization framework, **Topo4D**, to obtain high-fidelity 4D meshes and temporal-stable textures. **1)** To maintain the quality of photo-realistic rendering in 3DGS and also extract meshes and textures with **fixed topology**, we first explicitly link 3D Gaussians to the pre-defined topological facial geometry, named Gaussian Mesh. Hence, it strikes a balance between high expressive power and geometric structure fixedness to help learn topology-constrained meshes and photo-realistic textures. **2)** Considering that the topology may be destroyed during dynamic optimization, we design physical and geometric prior constraint terms on topological relationships during the optimization process to ensure **temporal topological stability** and regular mesh arrangement. Moreover, inheriting the advantages of 3DGS, the optimization process is computationally fast and memory-efficient. **3)** Once optimized, we align the Gaussian surface with the rendered surface using geometric normal prior to extracting high-quality meshes. In addition, we design the UV densification module to learn ultra-high-resolution textures with pore-level details by inversely mapping the Gaussian colors into UV space.

Our approach's effectiveness has been validated through extensive experiments. To our knowledge, Topo4D is the first method to implement the generation of high-fidelity 4D facial models with native 8K texture mapping, demonstrating the potential of Gaussian for dynamic face reconstruction and ultra-high-resolution tasks. In summary, our contributions include:

- We propose a novel optimization framework, Topo4D, for the reconstruction of high-quality 4D heads and photo-realistic textures with pore-level details from multi-view videos.
- We propose the Gaussian Mesh with UV densification to better represent facial models in the pre-defined topology and fixed UV.
- We design the alternative geometry and texture optimization process to ensure temporal topology stability and regular mesh arrangement during the optimization process.

2 Related Works

2.1 Registered Facial Model Acquisition

Acquiring high-fidelity facial models in pre-defined topologic structures has been a long-standing research challenge. Beginning with 3D Morphable Models [11], many methods [2–4,7,21,26,36,52,57,59,60] employ parametric models to achieve face reconstruction with single- or multi-view images. However, such methods struggle to faithfully reconstruct faces. Non-rigid ICP [9,13,15,29] methods deform canonical models to fit scans reconstructed by MVS methods with high quality. However, naively extending these methods to 4D videos will result in temporal instability, leading to texture drift. To maintain temporal stability, high-precision models are re-topology frame-by-frame in existing CG pipelines using professional software, *e.g.*, Wrap4D [1], which demands extensive time and expertise from experienced artists. To solve this, some methods [8,14,24] utilize optical flow or other techniques as supervision to deform the template models. However, numerous hyper-parameters need to be carefully tuned in these methods, making it difficult to generalize to different identities, and they are computationally slow. Another category of methods [10,12,37,43] directly regresses models from images. while these methods are constrained by a large number of training data and face difficulties in extending to other capture systems. Besides, out-of-domain expressions may be limited by insufficient training data. In this paper, we propose a novel approach for acquiring registered facial models with the quality typically achieved by artists manually, but in significantly less time.

2.2 Facial UV-Texture Recovery

Traditional CG pipelines predominantly use inverse rendering on reconstructed meshes to acquire textures from images at a computationally slow speed. To accelerate the process, many methods [4, 10, 20, 34, 35, 55, 56, 69] directly extract features from images to generate textures, where the quality is limited by resolution. Despite employing super-resolution networks [34, 43], they may still arise artifacts on the texture, and cannot accurately replicate pore-level details. Moreover, the aforementioned approaches are primarily designed for static tasks, and thus may not ensure the temporal stability of textures. Notably, Zhang et al. [69] can achieve video-level texture generation. However, it can only produce wrinkle maps that can be composited with natural high-resolution textures to represent varied expressions, rather than directly generating textures with high-frequency details, limiting its availability. Compared to these approaches, our method ensures temporal topological consistency in texture recovery and can directly generate textures in native 8K resolution with pore-level details.

2.3 Scene Representation

Neural Radiance Fields (NeRF) [49] has garnered significant attention for its remarkable capability to faithfully preserve both geometric and texture details

of objects. Subsequent advancements in training speed [16, 23, 50], inference speed [18,41,44], geometric quality [48,58,68], rendering quality [5,6,30], and dynamic scene representation [22, 25, 38, 64] have considerably enhanced the applicability of NeRF. Furthermore, 3D Gaussian splatting (3DGS) [31] has achieved SOTA results in scene representation due to its high-fidelity rendering, efficient training, and inference speeds, as well as memory efficiency. This technique has been further extended to 4D scene reconstruction [40, 46, 61, 66, 67], with notable applications in representing dynamic heads [17, 53, 63] and bodies [39, 54, 70]. However, these methods typically utilize continuous neural networks or random Gaussians to represent objects, posing challenges in extracting meshes with fixed topological structures, thus limiting their integration with existing industrial processes. To address these issues, we propose a meticulously improved 3DGS framework that can extract dynamic high-quality meshes and photo-realistic textures in constrained topology and UV from multi-view videos. Additionally, our method can be directly applied to current computer graphic industrial processes.

3 Methods

Our method aims to achieve temporally stable head mesh reconstruction and texture recovery from calibrated multi-view videos. Specifically, given sets of multi-view image sequences $\{\mathbf{I}_i^j \in \mathbb{R}^{h\times w\times 3} | 0 \leq i \leq F-1\}_{j=1}^{K}$ in the resolution of $h \times w$, encompassing F frames captured from K different viewpoints, all with known camera calibrations, our method can extract head meshes $\{\mathbf{S}_i := (V^i, T) | V^i \in \mathbb{R}^{n_v\times 3}\}_{i=0}^{F-1}$ in the pre-defined fixed topology T together with texture maps $\{\mathbf{M}_i \in \mathbb{R}^{8192\times 8192\times 3}\}_{i=0}^{F-1}$, where n_v represents the number of vertices.

To begin with, we give a brief review of 3D Gaussian Splatting [31] (Sec. 3.1). Our method first builds a *Gaussian Mesh* by initializing a topology-integrated Gaussian set based on the facial priors in the first frame (Sec. 3.2). Then, for each subsequent frame, we alternatively perform geometry optimization and texture optimization, to learn dynamic high-fidelity geometries and textures (Sec. 3.3). Finally, we introduce how to extract geometries from *Gaussian Mesh* and recover ultra-high-resolution textures (Sec. 3.4). The full pipeline is illustrated in Fig. 2.

3.1 Preliminary

3D Gaussian Splatting (3DGS) [31] is proposed as a competitive solution for photo-realistic rendering. Different from other implicit methods, 3DGS explicitly maintains a set of Gaussian distributions to model a scene. In 3DGS, each ellipsoidal Gaussian features a learnable color component $\boldsymbol{c}$ and an opacity component $\boldsymbol{\sigma}$, and is described by a covariance matrix Σ and its mean position $\boldsymbol{\mu}$:

$$G(\boldsymbol{x}) = e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}, \tag{1}$$

where Σ is further decomposed into rotation matrix R, parameterized with a quaternion $\boldsymbol{q}$, and scaling matrix S :

$$\Sigma = RSS^TR^T. \tag{2}$$

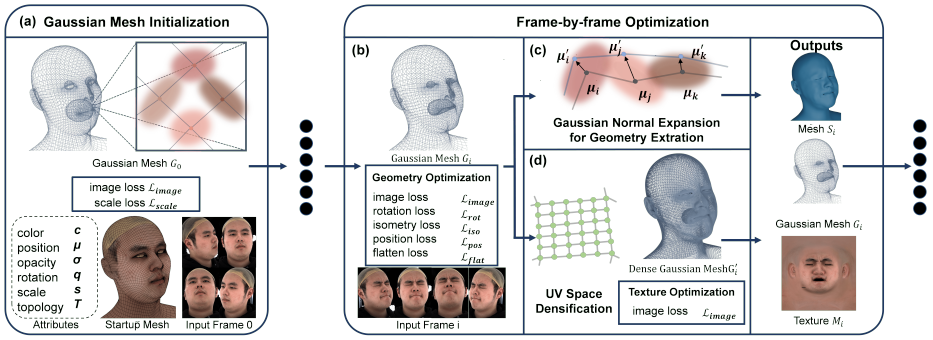

Fig. 2: Overall pipeline of our framework. **(a)** We initialize Gaussian attributes and establish topological correspondence with the startup mesh. **(b)** Take one frame as an example, geometry-related attributes in the Gaussian Mesh of the last frame are optimized by this frame under a set of topology-aware loss items. **(c)** We align the Gaussian surface with the rendering surface by Gaussian normal expansion to extract more precise meshes. **(d)** To learn pore-level detailed colors and generate ultra-high resolution texture, we build a dense mesh by densifying Gaussians in UV space.

In rendering, the color $\boldsymbol{C}$ of a pixel is acquired by sampling and blending all Gaussians that overlap the pixel in depth order:

$$\boldsymbol{C} = \sum_{i=1} \boldsymbol{c}_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{3}$$

where the blending weight α_i is given by evaluating a 2D Gaussian with its covariance multiplied by its opacity.

However, 3DGS is designed for realistic rendering instead of 3D reconstruction. Gaussians lack inherent topological relationships, thus unconstrained optimization of their attributes may only yield irregular geometry. As a result, directly extracting topologically sound meshes becomes unfeasible.

3.2 Gaussian Mesh for Topology Integrated Gaussians Initialization

Extracting topologically consistent meshes from Gaussians is challenging due to the lack of geometric constraints during optimization. To this end, we propose **Gaussian Mesh**, which uniquely incorporates the topological prior into vanilla Gaussian and will not affect its high-fidelity rendering quality. We define 4D Gaussian Meshes as $G_i = \{G_{i,j}\}_{j=1}^{n_v}$ for the i-th frame, where $G_{i,j}$ covers some learnable parameters, *i.e.*, $\{\boldsymbol{\mu}_{i,j} \in \mathbb{R}^3, \boldsymbol{q}_{i,j} \in \mathbb{R}^4, \boldsymbol{s}_{i,j} \in \mathbb{R}^3, \boldsymbol{c}_{i,j} \in \mathbb{R}^3, \boldsymbol{\sigma}_{i,j} \in \mathbb{R}\}$ for the position, rotation, scaling, color, and opacity separately of the j-th vertice in pre-defined topology T.

Different from 3DGS [31] using SFM-generated messy sparse points for initialization, to directly obtain pre-defined topological information, we first initialize Gaussian Mesh with head mesh and texture of the first frame, which is

acquired by automatic MVS and ICP algorithms. Specifically, we set the mean positions $\boldsymbol{\mu}_0$ of Gaussians to the corresponding 3D coordinates of vertices in topological order. Furthermore, to better align Gaussians with the surface, we initialize the orientation $\boldsymbol{q}_0$ of each Gaussian with the normal direction of vertices. It is worth noting that recent GaussianAvatars [53] also rigs Gaussians to face model. It aims at driving Gaussians with parametric models for photo-realistic rendering, where meshes are obtained by optimizing FLAME [36] parameters with the utilization of landmarks, personal displacements, and other additional supervision. In contrast, our goal is to extract high-quality topologic meshes and textures from multi-view videos by directly tracking the sequences without preprocessing registration and tracking.

After initializing the shape-related attributes ($\boldsymbol{\mu}_0$ and $\mathbf{q}_0$) of Gaussians, we then optimize their rendering-related attributes ($\boldsymbol{s}_0$, $\boldsymbol{c}_0$ and $\boldsymbol{\sigma}_0$). Concretely, we initialize the scales $\boldsymbol{s}_0$ as half of the minimum distance between each Gaussian and its one-ring neighbors and set opacity $\boldsymbol{\sigma}_0$ to 1, to avoid color blending between multiple Gaussians. To faithfully represent the color, we also initialize the color $\boldsymbol{c}_0$ of each Gaussian by sampling the corresponding pixel on the texture map according to the UV coordinate. Finally, to learn more details in dense parts, *e.g.*, eyes and mouth, we optimize Gaussian's rotation $\boldsymbol{q}_0$ and scale $\boldsymbol{s}_0$ between the first frames and rendered images $\mathbf{I}'_0$ with the same loss function as 3DGS:

$$\mathcal{L}_{image} = (1 - \lambda_{image})\mathcal{L}_1(\mathbf{I}_0, \mathbf{I}'_0) + \lambda_{image}\mathcal{L}_{D-SSIM}(\mathbf{I}_0, \mathbf{I}'_0), \tag{4}$$

Also, Gaussians have volume, leading to a certain gap between the center of Gaussians and the true surface. Therefore, the thickness of the Gaussian in the normal direction should be as small as possible. Therefore, we propose a scale loss to encourage the minimum value of every Gaussian's scale close to 0 and penalize Gaussians with a scale exceeding λ_{init} times than its initial value $\boldsymbol{s}_{init,i}$:

$$\mathcal{L}_{scale} = \sum_{i \in G}(\|\boldsymbol{s}_{0,i}\|_{-\infty} + max(0, \boldsymbol{s}_{0,i} - \lambda_{init}\boldsymbol{s}_{init,i})). \tag{5}$$

Overall, the final loss to initialize Gaussian Mesh can be formulated as:

$$\mathcal{L}_{init} = \mathcal{L}_{image} + \lambda_{scale}\mathcal{L}_{scale}. \tag{6}$$

3.3 Alternative Geometry and Texture Optimization

After initializing Gaussian Mesh G_0, we propose an **Alternative Geometry and Texture Optimization method** to acquire Gaussian Mesh geometry and learn dense texture colors. Specifically, at frame t, we first optimize Gaussians G_t by tracking G_{t-1} under the regularization of topology and physics. Afterward, dense texture color can be learned based on the tracked geometry. We perform such an alternative optimization process once per frame.

Geometry Optimization. Naively optimizing Gaussian Mesh causes topological confusion. To maintain the topology within Gaussians and regular mesh arrangement, we extend 3DGS [31] by introducing physical and topological priors.

Specifically, we propose physical loss item $\mathcal{L}_{phy}$ that constrain local rigidity and topological loss $\mathcal{L}_{topo}$ that improve the mesh quality. Overall, the loss functions for geometry optimization are constructed in three parts:

$$\mathcal{L}_{geo} = \mathcal{L}_{image} + \mathcal{L}_{phy} + \mathcal{L}_{topo}, \tag{7}$$

Physical Prior Loss. Solely utilizing color as supervision to optimize Gaussians is disastrous since low-frequency texture details, *e.g.*, forehead and cheek, lead to point mistracking, thus breaking the facial topology. To better regularize the motion of Gaussians and maintain topologic information, we modify the loss functions in Luiten et al. [46] with the constraints by one-ring neighbors as:

$$\mathcal{L}_{rot} = \frac{1}{2n_e} \sum_{i \in G} \sum_{j \in \mathcal{K}_i} w_{i,j} \| \hat{\boldsymbol{q}}_{t,j} \hat{\boldsymbol{q}}_{t-1,j}^{-1} - \hat{\boldsymbol{q}}_{t,i} \hat{\boldsymbol{q}}_{t-1,i}^{-1} \|_2, \tag{8}$$

where $\hat{\boldsymbol{q}}$ is the normalized quaternion, n_e is the number of edges, and $\mathcal{K}_i$ means the one-ring neighbours of $G_{t,i}$. The loss weighing factor w takes into account the edge length of the Gaussian Mesh at the first frame:

$$w_{i,j} = exp(-\lambda_w \| \boldsymbol{\mu}_{0,j} - \boldsymbol{\mu}_{0,i} \|_2^2). \tag{9}$$

In addition to the rotation similarity calculated between adjacent frames, we find that long-term physical loss is important for maintaining long-term stable dense correspondence:

$$\mathcal{L}_{iso} = \frac{1}{2n_e} \sum_{i \in G} \sum_{j \in \mathcal{K}_i} w_{i,j} | \| \boldsymbol{\mu}_{0,j} - \boldsymbol{\mu}_{0,i} \|_2 - \| \boldsymbol{\mu}_{t,j} - \boldsymbol{\mu}_{t,i} \|_2 |. \tag{10}$$

Finally, our physical loss items is the weighted sum of these two loss functions:

$$\mathcal{L}_{phy} = \lambda_{rot} \mathcal{L}_{rot} + \lambda_{iso} \mathcal{L}_{iso}. \tag{11}$$

Topology Prior Loss. Physical constraints achieve long-term tracking of the corresponding Gaussians, but they will lead to irregular wiring and unsmooth surfaces. To realize regular wiring, we calculate the L2 loss between the position of each vertex and the average position of its neighbors:

$$\mathcal{L}_{pos} = \frac{1}{n_v} \sum_{i \in G} (\boldsymbol{\mu}_{t,i} - \frac{\sum_{j \in \mathcal{K}_i} \boldsymbol{\mu}_{t,j}}{|\mathcal{K}_i|})^2. \tag{12}$$

To maintain the stable topology and smooth surface, we further apply mesh flattening loss to the angles between adjacent faces in optimization:

$$\mathcal{L}_{flat} = \sum_{\theta_i \in e_i} (1 - cos(\theta_{t,i} - \theta_{0,i})), \tag{13}$$

where $\theta_{t,i}$ is the angle between the faces that have the common edge e_i at frame t. Overall, our topological prior loss servers are as:

$$\mathcal{L}_{topo} = \lambda_{pos} \mathcal{L}_{pos} + \lambda_{flat} \mathcal{L}_{flat}. \tag{14}$$

Texture Optimization. After obtaining the geometry with the coarse texture, we continue to learn high-detailed dense texture. Compared to representing geometry, generating an **8K** texture map with pore-level details requires more Gaussians to learn from high-resolution images. Different from vanilla 3DGS, which adopts an adaptive densification process that results in a messy topology, we propose a novel densification method, *i.e.*, **UV Space Densification**, which allows for topologically densified Gaussians corresponding to the UV space of the texture map. Specifically, at frame t, we first initialize the dense Gaussian Mesh G'_t with the Gaussians in G_t. Then, we insert each quadrilateral grid in G'_t with $(N \times N)$ smaller grids by bi-linear interpolation, with each new Gaussian on each inserted vertex. The attributes and UV coordinates of these new Gaussians are also bi-linear interpolated with Gaussians in G_t. We optimize dense Gaussian Mesh G'_t by Eq. 4 and follow the same practice as Sec. 3.2 to learn high-frequency texture with sub-micron details.

3.4 Extracting Geometry and Texture from Gaussians

After completing the geometry and texture optimization, we can extract 4D meshes from $\{G_i\}_{i=0}^{F-1}$ and 8K texture maps from $\{G'_i\}_{i=0}^{F-1}$.

Geometry Extraction. Despite special regularization for the scale of Gaussian, it still has volume, resulting in a slight decrease in geometry reconstructed by directly extracting Gaussians' positions. Therefore, we propose Gaussian Normal Expansion to make surfaces derived from Gaussian surfaces resemble real surfaces more closely. As illustrated in Fig. 2 (c), we offset each Gaussian in the direction of the vertex normal by its projection scale in normal direction to obtain the final meshes $\{\mathbf{S}_i\}_{i=0}^{F-1}$.

Texture Extraction. To generate 8K texture maps from a dense Gaussian mesh $\{G'_i\}_{i=0}^{F-1}$, we map Gaussians' colors to UV space based on their UV coordinates. Since our Gaussian meshes are topologized, we can triangulate them and map learned dense texture colors to texture maps $\{\mathbf{M}_i\}_{i=0}^{F-1}$ by a rasterization-based forward rendering method [33].

4 Experiments

4.1 Dataset and Implementation Details

Data Preparation. We collect a dynamic multi-view head dataset using a Light Stage [19] with 16 calibrated color cameras. In the dataset, images are captured at a resolution of 4096 × 3000 and a rate of 60 fps. We capture multi-view videos for 10 identities. Each identity should perform an expression sequence and a talking sequence separately, with each expression sequence containing diverse expressions, including extreme and asymmetric ones. Each sequence lasts between 400 to 600 frames and is required to begin from **a neutral expression.**

Implementation Details. We implement our method based on PyTorch [51] and NVIDIA 3090 GPUs. The mesh topology consists of $n_v = 8280$ vertices and

Table 1: Quantitative evaluation for different face reconstruction methods on our prepared dataset. We measure the percentage of vertices within different error levels and calculate the mean error and variance.

Type	Methods	<0.2mm(%)↑	<0.5mm(%)↑	<1mm(%)↑	<2mm(%)↑	<3mm(%)↑	Mean(mm)↓	Med.(mm)↓
Single-view	DECA [21]	2.055	5.136	10.264	20.075	29.046	8.104	5.929
	HRN [10]	5.170	12.786	20.734	44.692	60.263	2.871	2.429
Multi-view	MVFR [62]	4.139	10.130	19.407	34.629	43.661	7.800	4.357
	DFNRMVS [3]	3.447	8.579	17.064	33.479	48.356	3.649	3.214
	Ours	**22.485**	**52.856**	**87.376**	**94.379**	**97.697**	**0.686**	**0.471**

$n_f = 16494$ faces. Since our Light Stage can provide uniform lighting, we use RGB rather than SH in 3DGS to represent view-consistent colors, which is more efficient and easier to optimize. We use Adam [32] for optimization. The geometry optimization stage includes 1000 iterations at each timestamp, with all images resized to 512×375. We mask out the inner mouth using a face parsing model [42] to prevent the vertices around the lips from learning incorrect colors. The texture optimization stage consists of 300 iterations at each timestamp and is learned at the original 4K resolution without preprocessing, with a dense number $N = 30$. We set $\lambda_{image} = 0.2$, $\lambda_{scale} = 10$, $\lambda_{rot} = 20$, $\lambda_{iso} = 20$, $\lambda_{pos} = 1e3$, and $\lambda_{flat} = 2e - 4$. We employ an MVS method [45] to reconstruct 4D head scans for evaluation and use an automatic iterative closest point (ICP) [9] algorithm to obtain a roughly accurate mesh in the first frame for initialization. The texture outside the facial area is obtained by alpha blending with the template textures. All hyper-parameters remain the **same** for all experiments. Please refer to the Supplementary Material for more details.

4.2 Face Reconstruction

Baseline. We evaluate our method on our collected dataset and compare it to three types of SOTA topology-consistent face reconstruction methods: (1) single-view methods DECA [21], and HRN [10]; (2) multi-view methods MVFR [62] and DFNRMVS [3]; (3) a traditional MVS and ICP pipeline w/wo facial landmark guidance. For single-view methods, we input the front-view image. For multi-view methods, we use all 16 views. Note that, we follow DFNRMVS's setting that uses a front view and an oblique side view image to produce the best results.
Quantitative Comparisons. We evaluate facial reconstruction accuracy using mesh-to-scan distances. To minimize inconsistencies among different models, we only compute metrics on the facial region, excluding the ears, back of the head, and neck. In Tab. 1, our method significantly outperforms other topology-consistent methods in all metrics. Notably, the majority of vertices are located in high-precision ranges, with 52.8% more vertices within 0.5 mm precision. This improvement is attributed to the introduction of reliable geometry and color priors during Gaussian initialization.
Qualitative Comparisons. Fig. 3 shows a visual comparison between meshes reconstructed using different methods. We manually register the faces as the

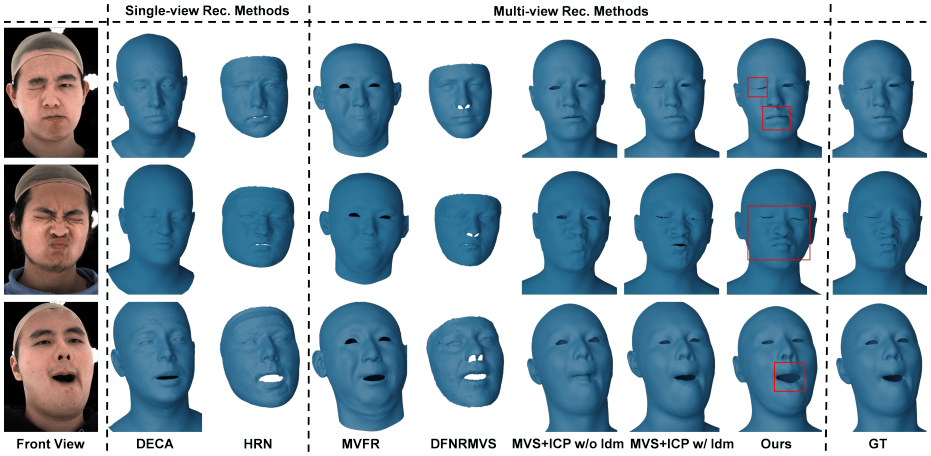

Fig. 3: Qualitative evaluation of meshes generated by our method and other topology-consistent reconstruction methods. We use artist-manually registered head mesh as the ground truth. We highlight areas that are difficult to reconstruct.

ground truth. Our method can faithfully capture both asymmetric extreme expressions and minor facial changes, outperforming other pre-trained methods. Additionally, we compare our approach with the traditional optimization-based MVS [45] + ICP [9] pipeline, guided by landmarks. Notably, even the advanced landmark detectors produce significant errors under extreme expressions, resulting in semantically incorrect registration, and fine areas are prone to serration and interpenetration. In contrast, our method reconstructs comparable details with correct correspondence.

4.3 Ultra-High Resolution Texture Generation

We qualitatively compare our generated textures with state-of-the-art facial texture estimation models: UnsupTex [56] and HRN [10]. Fig. 4 shows the rendering results of different methods, including the generated 8K textures. In the comparisons, UnsupTex and HRN are limited to producing low-resolution textures that lack realistic details. In contrast, our method directly generates high-quality 8K textures without up-sampling, faithfully capturing facial wrinkles, hair, and pores, and achieving noticeably superior rendering quality.

4.4 Temporal Stability Comparisons

We compare the temporal stability of the generated meshes and textures with other methods. For meshes, We measure geometry consistency by calculating the RMSE between adjacent frames. Fig. 5 (a) shows the curve of each method over an expression sequence. Our method achieves much more temporally stable vertex tracking with extreme expressions, whereas other methods exhibit considerable performance fluctuations. We notice that, in some cases, the inter-frame

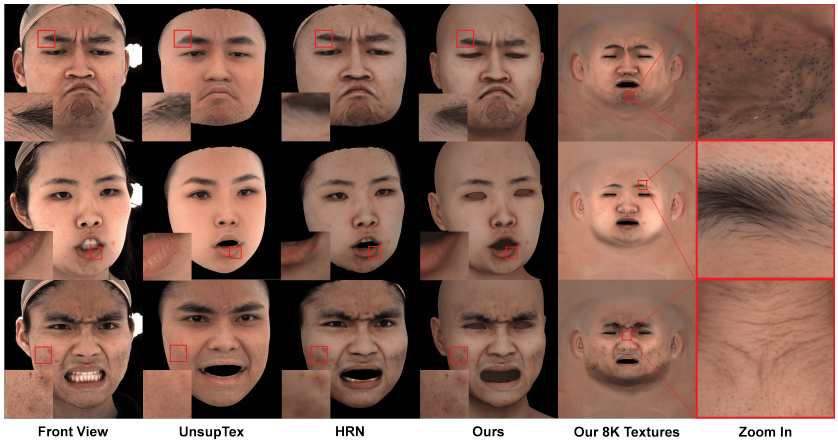

Fig. 4: Qualitative evaluation of the rendering results between our method, UnsupTex [56] and HRN [10]. The generated 8K textures and pore-level details are demonstrated in columns 5 and 6.

difference of DECA [21] is lower than ours. This is because DECA fails to capture some extreme expressions or minor facial changes, as shown in Fig. 3, and therefore tends to keep the meshes unchanged. Instead, our method can faithfully capture facial motions while maintaining temporal consistency. From the perspective of texture, we measure stability by comparing the PSNR of textures between consecutive frames. As shown in Fig. 5 (b), our method has the highest mean and lowest variance, indicating superior temporal stability. In conclusion, our method demonstrates better temporal consistency compared to other methods in both geometry and texture.

4.5 Ablation Studies

In this section, we assess the impacts of the crucial designs and parameters used in our method. For geometry reconstruction, we ablate all loss items, the number of input views, and the Gaussian normal expansion operation. Regarding texture estimation, we explore the impact of UV densification density on texture quality, along with the influence of $\mathcal{L}_{scale}$.

Analysis of Loss Items. We remove each loss function and keep other settings constant in the ablation study. Fig. 6 shows the qualitative results. 3DGS's [31] powerful rendering ability allows Gaussians to render similar images with entirely different geometry. We find that it's essential to physically constrain Gaussians' rotation and distance to achieve correct dense correspondence, especially in obstructed areas and areas with dense vertices like lips and nostrils. Although constraining the Gaussian's scale seems to have a limited impact on geometry accuracy (an increase of around 0.1mm), it is crucial for maintaining pore-level texture details and avoiding blurring, as shown in Fig. 7. The topology prior losses $\mathcal{L}_{pos}$ and $\mathcal{L}_{flat}$ effectively maintain the stability of invisible areas, such as

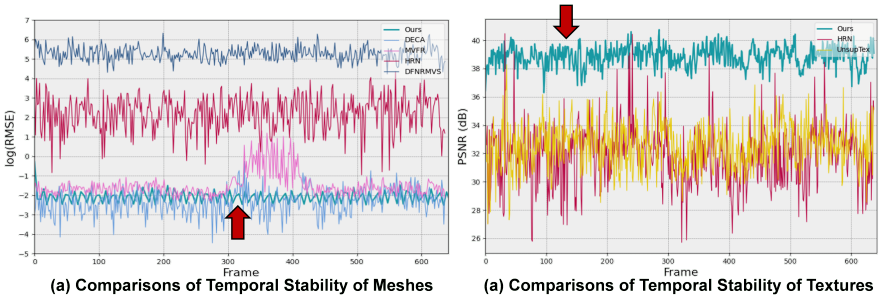

Fig. 5: Comparisons of temporal stability on a sequence in our dataset, and our method is bolded and indicated by red arrows. **(a)** The curves of log(RMSE) (lower is better) of several topology-consistent face reconstruction methods. **(b)** The curves of PSNR (higher is better) are calculated between textures of adjacent frames.

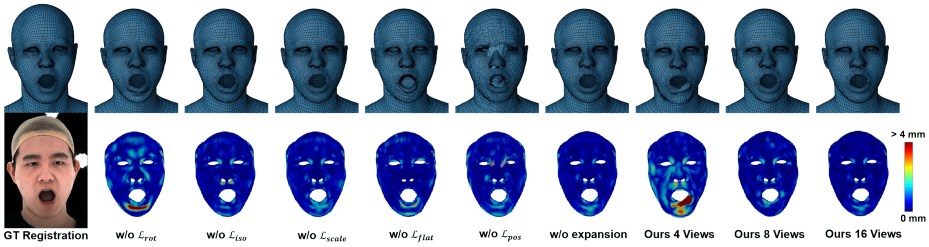

Fig. 6: Visualization of the reconstructed mesh after ablating loss items, Gaussian normal expansion operation, and view numbers. The second row displays the color-coded point-to-surface distance between the reconstructed mesh and the scan as a heatmap on the mesh's surface. **Please zoom-in for detailed observation.**

inner sockets, and significantly prevent exaggerated extrapolation or interpenetration. Ablating these mesh smoothing terms may result in lower mesh-to-scan errors, but it compromises the quality of the grid and wiring.

Analysis of Normal Expansion. Since we constrain the orientation and scale during initialization and optimization, Gaussians should display a small scale in the surface normal direction. As illustrated in Fig. 6, by using Gaussian normal expansion, we can marginally reduce the overall error.

Analysis of Input Views. We evaluate the effect of the number of input views on mesh quality, as shown in Fig. 6. It's obvious that even with half of the input views, our method can still yield competitive results, demonstrating its applicability to capture systems with fewer cameras. However, when the view number decreases to 4, the reconstruction mesh exhibits significant distortion.

Analysis of Gaussian Density. Fig. 7 depicts the impact of different density levels on texture quality in UV space densification. As the density decreases, the texture becomes blurred and loses detail. Our method can generate textures of any resolution directly. However, higher resolutions require denser Gaussians, more memory overhead, and longer optimization time.

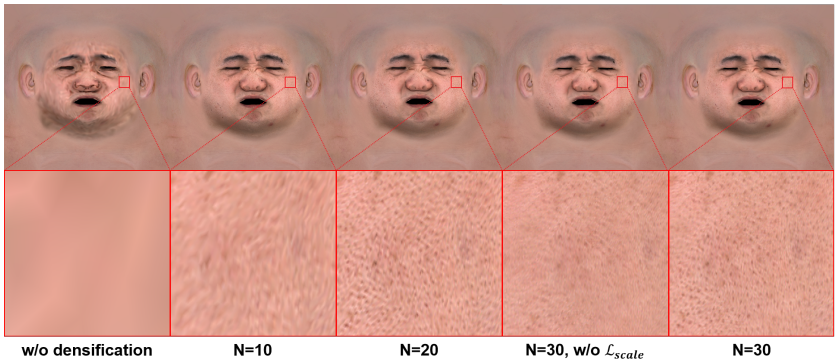

Fig. 7: Comparisons of texture quality under different settings.

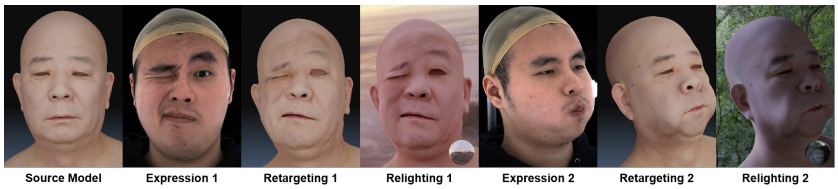

Fig. 8: We showcase the results of our method applied to retargeting the other face model(columns 3 and 6) and relighting results (columns 4 and 7).

4.6 Application

Driving digital characters by real performance is widely applied in industry workflows. Our method can retarget the captured expression sequence of an actor faithfully to other characters. Besides, relighting is also a widespread application in the CV and CG pipelines. Fig. 8 shows the results of retargeting extreme expressions and wrinkled texture maps learned by our methods to another facial model, as well as relighting rendering results.

5 Conclusion

In this paper, we propose Topo4D, an efficient framework that can extract temporal topology-consistent meshes and 8K textures from calibrated multi-view videos. Under the regularization of a set of topology-aware geometrical and physical loss items, we achieve topology-preserving Gaussian optimization while faithfully capturing the subject's expressions. By densifying Gaussians in UV space, we learn realistic pore-level details at high resolution and extract high-fidelity 8K texture maps. To sum up, our method provides a brand new way to reconstruct high-fidelity facial meshes and 8K texture maps, opening up new avenues for capturing 4D digital humans in an efficient and low-cost manner.

Acknowledgements

This work was supported in part by NSFC (62201342, 62101325), and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

References

1. Wrap4d - faceform. https://faceform.com/wrap4d/ 4
2. Aldrian, O., Smith, W.A.: Inverse rendering of faces with a 3d morphable model. TPAMI pp. 1080–1093 (2012) 2, 4
3. Bai, Z., Cui, Z., Rahim, J.A., Liu, X., Tan, P.: Deep facial non-rigid multi-view stereo. In: CVPR. pp. 5850–5860 (2020) 2, 4, 10
4. Bao, L., Lin, X., Chen, Y., Zhang, H., Wang, S., Zhe, X., Kang, D., Huang, H., Jiang, X., Wang, J., Yu, D., Zhang, Z.: High-fidelity 3d digital human head creation from rgb-d selfies. TOG (2021) 2, 4
5. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: CVPR. pp. 5855–5864 (2021) 5
6. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: CVPR. pp. 5470–5479 (2022) 5
7. Bas, A., Smith, W.A., Bolkart, T., Wuhrer, S.: Fitting a 3d morphable model to edges: A comparison between hard and soft correspondences. In: Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13. pp. 377–391 (2017) 2, 4
8. Beeler, T., Hahn, F., Bradley, D., Bickel, B., Beardsley, P.A., Gotsman, C., Sumner, R.W., Gross, M.H.: High-quality passive facial performance capture using anchor frames. ACM Trans. Graph. p. 75 (2011) 2, 4
9. Besl, P.J., McKay, N.D.: Method for registration of 3-d shapes. In: Sensor fusion IV: control paradigms and data structures. pp. 586–606 (1992) 2, 4, 10, 11
10. Biwen, L., Jianqiang, R., Mengyang, F., Miaomiao, C., Xuansong, X.: A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images. In: CVPR. pp. 394–403 (2023) 4, 10, 11, 12
11. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: SIGGRAPH, pp. 187–194 (1999) 4
12. Bolkart, T., Li, T., Black, M.J.: Instant multi-view head capture through learnable registration. In: CVPR. pp. 768–779 (2023) 2, 4
13. Booth, J., Roussos, A., Zafeiriou, S., Ponniah, A., Dunaway, D.: A 3d morphable model learnt from 10,000 faces. In: CVPR. pp. 5543–5552 (2016) 4
14. Bradley, D., Heidrich, W., Popa, T., Sheffer, A.: High resolution passive facial performance capture. In: ACM SIGGRAPH. pp. 1–10 (2010) 2, 4
15. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: A 3d facial expression database for visual computing. TVCG pp. 413–425 (2013) 4
16. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: ECCV. pp. 333–350 (2022) 5
17. Chen, Y., Wang, L., Li, Q., Xiao, H., Zhang, S., Yao, H., Liu, Y.: Monogaussianavatar: Monocular gaussian point-based head avatar. arXiv preprint arXiv:2312.04558 (2023) 5

18. Chen, Z., Funkhouser, T., Hedman, P., Tagliasacchi, A.: Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. In: CVPR. pp. 16569–16578 (2023) 5
19. Debevec, P.: The light stages and their applications to photoreal digital actors. SIGGRAPH Asia pp. 1–6 (2012) 2, 9
20. Deng, J., Cheng, S., Xue, N., Zhou, Y., Zafeiriou, S.: Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In: CVPR (2018) 4
21. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3D face model from in-the-wild images. In: TOG (2021) 2, 4, 10, 12
22. Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-planes: Explicit radiance fields in space, time, and appearance. In: CVPR. pp. 12479–12488 (2023) 5
23. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: CVPR. pp. 5501–5510 (2022) 5
24. Fyffe, G., Nagano, K., Huynh, L., Saito, S., Busch, J., Jones, A., Li, H., Debevec, P.: Multi-view stereo on consistent face topology. In: Computer Graphics Forum. pp. 295–309 (2017) 2, 4
25. Gao, C., Saraf, A., Kopf, J., Huang, J.B.: Dynamic view synthesis from dynamic monocular video. In: CVPR. pp. 5712–5721 (2021) 5
26. Gecer, B., Ploumpis, S., Kotsia, I., Zafeiriou, S.: Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In: CVPR (2019) 2, 4
27. Goesele, M., Curless, B., Seitz, S.M.: Multi-view stereo revisited. In: CVPR. pp. 2402–2409 (2006) 2
28. Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3d dense face alignment. In: ECCV (2020) 2
29. Ji, P., Li, H., Jiang, L., Liu, X.: Light-weight multi-view topology consistent facial geometry and reflectance capture. In: CGI. pp. 139–150 (2021) 4
30. Jiang, Y., Hedman, P., Mildenhall, B., Xu, D., Barron, J.T., Wang, Z., Xue, T.: Alignerf: High-fidelity neural radiance fields via alignment-aware training. In: CVPR. pp. 46–55 (2023) 5
31. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. TOG (2023) 3, 5, 6, 7, 12
32. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 10
33. Laine, S., Hellsten, J., Karras, T., Seol, Y., Lehtinen, J., Aila, T.: Modular primitives for high-performance differentiable rendering. TOG (2020) 9
34. Lattas, A., Moschoglou, S., Gecer, B., Ploumpis, S., Triantafyllou, V., Ghosh, A., Zafeiriou, S.: AvatarMe: Realistically renderable 3D facial reconstruction. In: CVPR (2020) 2, 4
35. Lattas, A., Moschoglou, S., Ploumpis, S., Gecer, B., Deng, J., Zafeiriou, S.: Fitme: Deep photorealistic 3d morphable model avatars. In: CVPR. pp. 8629–8640 (2023) 4
36. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. TOG pp. 194–1 (2017) 2, 4, 7
37. Li, T., Liu, S., Bolkart, T., Liu, J., Li, H., Zhao, Y.: Topologically consistent multi-view face inference using volumetric sampling. In: CVPR. pp. 3824–3834 (2021) 2, 4
38. Li, T., Slavcheva, M., Zollhoefer, M., Green, S., Lassner, C., Kim, C., Schmidt, T., Lovegrove, S., Goesele, M., Newcombe, R., et al.: Neural 3d video synthesis from multi-view video. In: CVPR. pp. 5521–5531 (2022) 5

39. Li, Z., Zheng, Z., Wang, L., Liu, Y.: Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. arXiv preprint arXiv:2311.16096 (2023) 5
40. Liang, Y., Khan, N., Li, Z., Nguyen-Phuoc, T., Lanman, D., Tompkin, J., Xiao, L.: Gaufre: Gaussian deformation fields for real-time dynamic novel view synthesis. arXiv preprint arXiv:2312.11458 (2023) 3, 5
41. Lin, H., Peng, S., Xu, Z., Yan, Y., Shuai, Q., Bao, H., Zhou, X.: Efficient neural radiance fields for interactive free-viewpoint video. In: SIGGRAPH Asia. pp. 1–9 (2022) 5
42. Lin, Y., Shen, J., Wang, Y., Pantic, M.: Roi tanh-polar transformer network for face parsing in the wild. Image and Vision Computing p. 104190 (2021) 10
43. Liu, S., Cai, Y., Chen, H., Zhou, Y., Zhao, Y.: Rapid face asset acquisition with recurrent feature alignment. TOG pp. 1–17 (2022) 2, 4
44. Lombardi, S., Simon, T., Schwartz, G., Zollhoefer, M., Sheikh, Y., Saragih, J.: Mixture of volumetric primitives for efficient neural rendering. TOG pp. 1–13 (2021) 5
45. Loop, C., Zhang, Z.: Computing rectifying homographies for stereo vision. In: CVPR. pp. 125–131 (1999) 10, 11
46. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. arXiv preprint arXiv:2308.09713 (2023) 3, 5, 8
47. Ma, W.C., Hawkins, T., Peers, P., Chabert, C.F., Weiss, M., Debevec, P.E., et al.: Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. Rendering Techniques p. 10 (2007) 2
48. Meng, X., Chen, W., Yang, B.: Neat: Learning neural implicit surfaces with arbitrary topologies from multi-view images. In: CVPR. pp. 248–258 (2023) 5
49. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM pp. 99–106 (2021) 4
50. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. TOG pp. 1–15 (2022) 5
51. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIPS **32** (2019) 9
52. Ploumpis, S., Ververas, E., O'Sullivan, E., Moschoglou, S., Wang, H., Pears, N., Smith, W.A., Gecer, B., Zafeiriou, S.: Towards a complete 3d morphable model of the human head. TPAMI pp. 4142–4160 (2020) 2, 4
53. Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: Gaus sianavatars: Photorealistic head avatars with rigged 3d gaussians. arXiv preprint arXiv:2312.02069 (2023) 5, 7
54. Qian, Z., Wang, S., Mihajlovic, M., Geiger, A., Tang, S.: 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. arXiv preprint arXiv:2312.09228 (2023) 5
55. Ren, X., Lattas, A., Gecer, B., Deng, J., Ma, C., Yang, X.: Facial geometric detail recovery via implicit representation. In: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–8. IEEE (2023) 4
56. Slossberg, R., Jubran, I., Kimmel, R.: Unsupervised high-fidelity facial texture generation and reconstruction. In: ECCV (2022) 4, 11, 12
57. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: CVPR. pp. 2387–2395 (2016) 2, 4

58. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. NeurIPS pp. 27171–27183 (2021) 5
59. Wood, E., Baltrušaitis, T., Hewitt, C., Johnson, M., Shen, J., Milosavljević, N., Wilde, D., Garbin, S., Sharp, T., Stojiljković, I., et al.: 3d face reconstruction with dense landmarks. In: ECCV. pp. 160–177 (2022) 2, 4
60. Wu, F., Bao, L., Chen, Y., Ling, Y., Song, Y., Li, S., Ngan, K.N., Liu, W.: Mvf-net: Multi-view 3d face morphable model regression. In: CVPR. pp. 959–968 (2019) 2, 4
61. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. arXiv preprint arXiv:2310.08528 (2023) 3, 5
62. Xiao, Y., Zhu, H., Yang, H., Diao, Z., Lu, X., Cao, X.: Detailed facial geometry recovery from multi-view images by learning an implicit function. In: AAAI (2022) 2, 10
63. Xu, Y., Chen, B., Li, Z., Zhang, H., Wang, L., Zheng, Z., Liu, Y.: Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. arXiv preprint arXiv:2312.03029 (2023) 5
64. Yan, Y., Zhou, Z., Wang, Z., Gao, J., Yang, X.: Dialoguenerf: Towards realistic avatar face-to-face conversation video generation. arXiv preprint arXiv:2203.07931 (2022) 5
65. Yang, H., Zhu, H., Wang, Y., Huang, M., Shen, Q., Yang, R., Cao, X.: Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In: CVPR. pp. 601–610 (2020) 2
66. Yang, Z., Yang, H., Pan, Z., Zhu, X., Zhang, L.: Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. arXiv preprint arXiv:2310.10642 (2023) 3, 5
67. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. arXiv preprint arXiv:2309.13101 (2023) 3, 5
68. Yariv, L., Hedman, P., Reiser, C., Verbin, D., Srinivasan, P.P., Szeliski, R., Barron, J.T., Mildenhall, B.: Bakedsdf: Meshing neural sdfs for real-time view synthesis. arXiv preprint arXiv:2302.14859 (2023) 5
69. Zhang, L., Zeng, C., Zhang, Q., Lin, H., Cao, R., Yang, W., Xu, L., Yu, J.: Video-driven neural physically-based facial asset for production. TOG pp. 1–16 (2022) 4
70. Zielonka, W., Bagautdinov, T., Saito, S., Zollhöfer, M., Thies, J., Romero, J.: Drivable 3d gaussian avatars. arXiv preprint arXiv:2311.08581 (2023) 5