TGAvatar: Reconstructing 3D Gaussian Avatars with Transformer-based Tri-plane

Ruigang Hu, Xuekuan Wang, Yichao Yan and Cairong Zhao, *Member, IEEE,*

***Abstract*—We introduce TGAvatar, a novel framework for 3D head animation and reconstruction that revolutionizes the use of 3D Gaussian Splatting (3DGS). TGAvatar significantly advances rendering quality by leveraging the intricate properties of 3DGS to achieve detailed and realistic representations of human head geometries and textures. We use an innovative application of linear blending techniques to imitate 3D Morphable Model (3DMM) coefficients within 3DGS, thereby enabling precise and dynamic facial feature and expression modeling. Further enhancing TGAvatar's capabilities, a transformer based tri-plane module is incorporated to accurately infer spherical harmonics and alpha parameters. This integration is pivotal for the method, as it allows allows us to efficiently and precisely represent the visual characteristics of gaussians, tailored specifically to the intricate details of the head's components. Our exhaustive evaluations show that TGAvatar not only elevates the fidelity and realism of 3D head reconstructions but also sets a new standard by surpassing existing methods in rendering quality and computational efficiency. Please see our project page at https://hrg0417.github.io/TGAvatar/**

***Index Terms*—3D Gaussian Splatting, facial animation, Computer Vision, Deep Learning.**

I. Introduction

THE digital recreation of human appearance, particularly the accurate and lifelike rendering of human heads, remains one of the most challenging and sought-after objectives in computer graphics, virtual reality, and augmented reality. This endeavor, crucial for creating immersive virtual experiences, telepresence, and digital entertainment, has seen considerable advancements over the years. The advent of 3D Morphable Models (3DMMs) marked a significant milestone by providing a framework for synthesizing facial geometries and expressions from a dataset of scanned 3D faces [2]. These models have been instrumental in pioneering the field, enabling the generation of new faces through the manipulation of a set of parameters controlling identity, expression, and other facial attributes [3], [4].

However, as the demand for more realistic and dynamically adaptable avatars grows, the limitations of traditional approaches become increasingly apparent. While 3DMMs excel in capturing a wide range of facial variations, their reliance on linear combinations of base models restricts their ability to represent finer details and textures, particularly in real-time scenarios or under varied lighting conditions. The emergence of Neural Radiance Fields (NeRF) [5] and its subsequent adaptations [6]–[8] introduced groundbreaking methods for rendering complex scenes with unprecedented detail and realism. Following works expanded NeRF to the field of high-definition facial characters and animation [9], [10]. These techniques, leveraging the power of deep learning, have set new standards for photorealistic rendering. Nonetheless, the computational intensity required for training and rendering, along with challenges in real-time adaptability and dynamic content generation, highlights the need for innovative solutions that balance fidelity, efficiency, and flexibility.

The development of 3D Gaussian Splatting (3DGS) [1] has significantly boosted the efficiency of rendering in novel view synthesis. In contrast to the neural implicit methods such as Neural Radiance Fields (NeRF) [5], which capture a 3D scene using position and viewpoint-dependent neural networks, 3D Gaussian Splatting adopts Gaussian ellipsoids as the modeling basis. This shift enables quicker rendering because these ellipsoids can be directly transformed into images through rasterization. Some follow-up works has extended 3DGS, enabling it to render dynamic scenes containing temporal sequences [11]–[14]. Notably, several advancements have also emerged in the field of facial animation, leveraging the 3DGS framework to achieve high-fidelity and real-time facial expressions and movements [15]–[18].

While existing 3D Gaussian Splatting (3DGS) and facial animation methods have made significant strides in rendering dynamic scenes and capturing intricate details, they often struggle with flexibility in representing subtle facial expressions. These methods typically employ an MLP module to infer the parameter offsets for the Gaussian components [15], [17], which can limit the model's ability to capture the full spectrum of facial nuances. Moreover, existing approaches [17], [18] treat the alpha and spherical harmonic parameters the same as other parameters, which may result in a lack of flexibility in rendering subtle facial features, particularly around the mouth and eyes. GaussianHead [15] utilized a tri-plane module [19] to address this issue; however, the parameters in their tri-plane module are not directly related to the expression coefficients, which may lead to artifacts across various expressions.

To address the issues raised above, we propose TGAvatar, a novel framework that revolutionizes 3D head animation and reconstruction by integrating the strengths of 3D Gaussian Splatting (3DGS) [1] with advanced modeling techniques. By employing a linear blending technique that imitates the coefficient manipulation inherent in 3DMMs, TGAvatar achieves

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.

Manuscript received April 19, 2021; revised August 16, 2021.

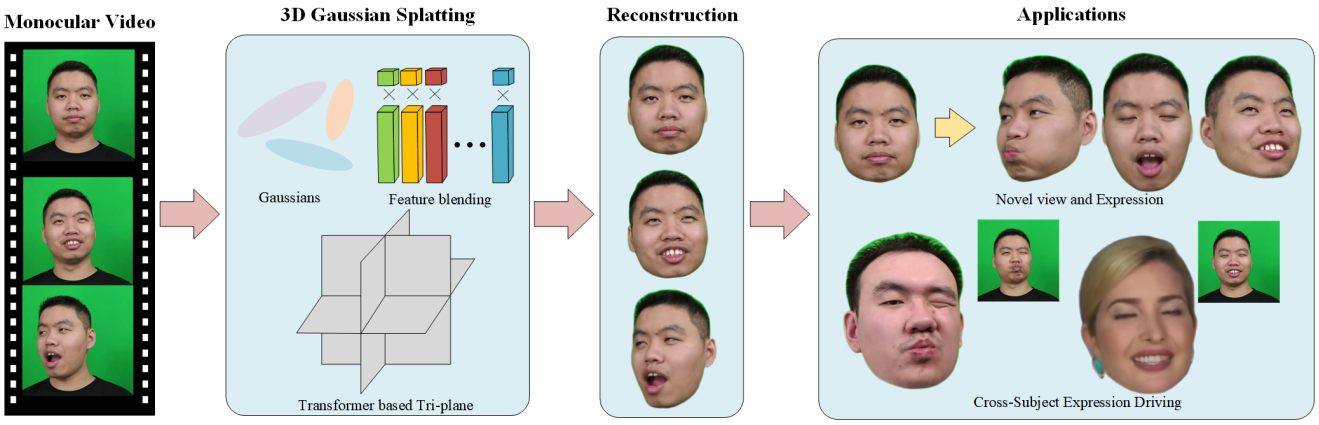

Fig. 1. TGAvatar reconstructs a 3D facial avatar from a monocular portrait video of a person. By leveraging 3D Gaussian Splatting [1], alongside 3DMM feature blending and a transformer based tri-plane module, TGAvatar gan generate lifelike novel views and expressions of the digital avatar.

a dynamic and precise modeling of facial features and expressions, , significantly enhancing the rendering quality of the head avatars.

Moreover, the incorporation of a tri-plane module [19] in TGAvatar marks a pivotal advancement, enabling the accurate inference of spherical harmonics and alpha parameters. Notably, our tri-plane module utilizes a transformer architecture conditioned on facial expressions, allowing it to effectively capture facial details specific to different expressions. This integration is critical for rendering photorealistic avatars under a variety of lighting conditions, addressing one of the most significant challenges in digital human representation. The tri-plane module, which has demonstrated its efficiency and expressive ability in past works [15], [20], [21], allows for a nuanced understanding of lighting and shading, essential for achieving lifelike renderings.

Our exhaustive evaluations and comparisons with existing methods underscore TGAvatar's superiority in rendering quality, computational efficiency, and the fidelity of 3D head reconstructions. TGAvatar not only represents a significant leap forward in the quest for realistic digital humans but also establishes a new benchmark for the field. The implications of our work are vast, promising to impact a wide range of applications, from virtual and augmented reality to digital filmmaking and beyond, where the realistic representation of human characters is paramount.

In summary, this paper makes the following three main contributions:

1) We introduced a novel method that integrates the principles of 3D Gaussian Splatting with techniques akin to those found in 3D Morphable Models, enabling the rapid generation of dynamic and realistic animations of facial expressions and movements in real time.

2) We developed a sophisticated tri-plane module derived from a transformer architecture, which significantly enhances the lifelike quality and detail of digital faces. This innovation allows our avatars to exhibit subtle shading and textural variations, resulting in a more natural and realistic appearance.

3) We tested our method on multiple datasets, and our results significantly outperform those reported in other studies, achieving state-of-the-art performance in terms of detail and efficiency for 3D reconstructions of human heads.

II. Related Work

3D Head Avatar Reconstruction: The quest for accurate 3D facial reconstruction has significantly evolved from its inception. Early efforts focused on geometric and photometric methods, leading to the development of 3D Morphable Models (3DMMs) by Blanz and Vetter [2], which offered a breakthrough in generating facial geometries by linearly combining a set of pre-scanned facial templates. While 3DMMs have been instrumental in pioneering facial animation and reconstruction, they often fall short in capturing high-fidelity details and dynamic expressions, particularly under diverse lighting conditions and extreme poses [3], [22], [23].

The limitations of traditional modeling techniques have spurred interest in leveraging machine learning for facial reconstruction. Deep learning approaches [24]–[27] have shown promise in capturing complex facial details beyond the capabilities of 3DMMs. These methods utilize convolutional neural networks to directly infer 3D facial structure from 2D images, offering improved flexibility and detail.

NeRF based facial avatar: Recent advancements in scene representation have been marked by the introduction of Neural Radiance Fields (NeRF) by Mildenhall et al. [5]. NeRF's ability to synthesize photorealistic images from sparse input data has revolutionized the field, leading to its application in dynamic facial animation and reconstruction [9], [10], [28]–[37]. Gafni et al. [9] introduced dynamic neural radiance fields tailored for monocular 4D facial avatar reconstruction, which captures the dynamic nuances of human faces, allowing for the synthesis of novel head poses and expressions directly from monocular video data. Park et al. [10] and Athar et al. [30] leveraged casual photos or videos to create deformable NeRF models, enabling photorealistic renderings of scenes with non-rigid transformations. Athar et al. [28] and Gao et al. [29] blended Neural Radiance Fields (NeRF) with 3D morphable models (3DMMs) for creating controllable, photorealistic portrait videos, which enabled novel view synthesis and facial expression manipulation using a low-dimensional expression

space. Guo et al. [31] utilized audio features to condition a dynamic neural radiance field, enabling the synthesis of photo-realistic talking-head videos with volume rendering. However, in order to render high-quality facial images, the complexity of neural networks is relatively high, leading to increased learning difficulty and computational demands.

To address these challenges, recent works have introduced hybrid neural fields that leverage both implicit and explicit data structures for scene representation. Explicit data structures such as tri-planes [19], hex-planes [38], [39], and voxels [40] are employed to alleviate the computational pressure on neural networks, allowing these networks to focus more on decoding semantic and intricate details of the scene. This hybrid approach enables a more efficient rendering process, particularly for complex scenes. Taken the advantages of these hybrid methods, many studies related to facial avatars have emerged [20], [41], [42]. Next3D [20] leveraged a novel representation called Generative Texture-Rasterized Tri-planes, which integrates fine-grained expression control of explicit mesh-driven deformation with the flexibility of implicit volumetric representation for animatable portrait synthesis. OTAvatar [41] proposed a method allows for the generation of face avatars from just a single portrait image using a tri-plane formulated volume rendering technique. Their core innovation is the decoupling-by-inverting strategy which separates identity and motion in the latent code via optimization-based inversion. HAvatar [42] introduced a parametric model-conditioned NeRF for personalized 3D head avatar creation, while a hybrid representation is developed that handles the inconsistent shape issue prevalent in existing NeRF-based avatar modeling methods, significantly enhancing animation stability.

While Neural Radiance Fields (NeRF) have made notable strides in photorealistic facial avatar reconstruction, they face significant challenges. NeRF-based methods are computationally intensive, often requiring substantial training time and high-capacity neural networks to manage detailed renderings and dynamic expressions. This makes real-time applications challenging due to increased latency.

3D Gaussian Splatting: The exploration of 3D Gaussian Splatting (3DGS) by Kerbl et al. [1] offers a novel solution, balancing the need for detail fidelity with computational efficiency, which overcomes the limitations of axis-aligned mappings, significantly enhancing performance in complex regions. The original 3DGS can only be used to render static scenes, afterwards, some approaches [11]–[14] have expanded the use of Gaussian representation for dynamic scene reconstruction. However, these methods cannot be directly applied to the reconstruction of facial avatars.

Recently, several subsequent studies [15]–[18] have further explored the area of facial avatars based on 3DGS. Mono-GaussianAvatar [16] leveraged a Gaussian deformation field for animating head avatars, allowing for adaptable topology and efficient rendering. The method demonstrated superior performance in creating photorealistic avatars, maintaining geometry stability, and effectively handling dynamic poses and expressions. FlashAvatar [17] combined Gaussian splatting with a 3D parametric face model, allowing for efficient reconstruction and rapid rendering. This method provided high visual quality with low computational costs, improving rendering speed and efficiency over previous approaches. GaussianBlendshapes [18] presented a Gaussian blendshape representation which consists of a base neutral model and expression blendshapes, all represented as 3D Gaussians. By learning these blendshapes from monocular video input, the approach is able to achieve realistic animations with high-frequency details, outperforming existing NeRF and point-based methods in terms of speed and quality. GaussianHead [15] employed a motion deformation field to adapt to facial ovements and a multi-resolution tri-plane to store appearance information of the head.

GaussianBlendshapes [18] utilizes a 3DMM feature blending structure, similar to our approach. However, in their approach, each blendshape corresponds to a separate Gaussian, while our approach integrates feature blending within a single Gaussian, and infer alpha and spherical harmonic (SH) via a tri-plane module. Additionally, although GaussianHead [15] also employs a tri-plane, it does not establish a connection with expression coefficients. In our method, the tri-plane utilizes expression coefficients as conditions, inferred through a transformer, enhancing the representation of dynamic facial attributes.

III. Method

In this section, we introduce our method for creating high-fidelity, dynamic 3D head avatars using 3D Gaussian Splatting (3DGS). Inspired by 3DMM, our approach employs feature blending techniques within each Gaussian to determine pose, rotation, and scale coefficients. To achieve superior rendering results, we first use a transformer-based tri-plane decoder to predict tri-plane features. Subsequently, we incorporate a tri-plane module to extract hybrid features based on the pose of each Gaussian. Finally, these hybrid features are fed into an MLP network to infer opacity and spherical harmonics coefficients.

In the following subsections, we introduce the preliminaries of tri-plane and Gaussian Splatting in Section III-A, describe the feature blending techniques in Section III-B, detail the hybrid tri-plane representation in Section III-C, and outline the training details and objectives in Section III-D.

A. Preliminary of Tri-plane and Gaussian Splatting

Tri-plane Hybrid 3D Representation: The tri-plane representation aligns explicit features along three orthogonal planes—XY, XZ, and YZ—with each plane having a resolution of $N \times N \times C$, where N is the spatial resolution and C is the number of channels. Any 3D position $x \in \mathbb{R}^3$ is projected onto these planes, retrieving the corresponding feature vectors F_{xy}, F_{xz}, F_{yz} via bilinear interpolation. These vectors are then summed to form a single feature vector for the position. A lightweight decoder, typically a small MLP, interprets the aggregated features to output color and density, which are used for neural volume rendering to produce RGB images [19].

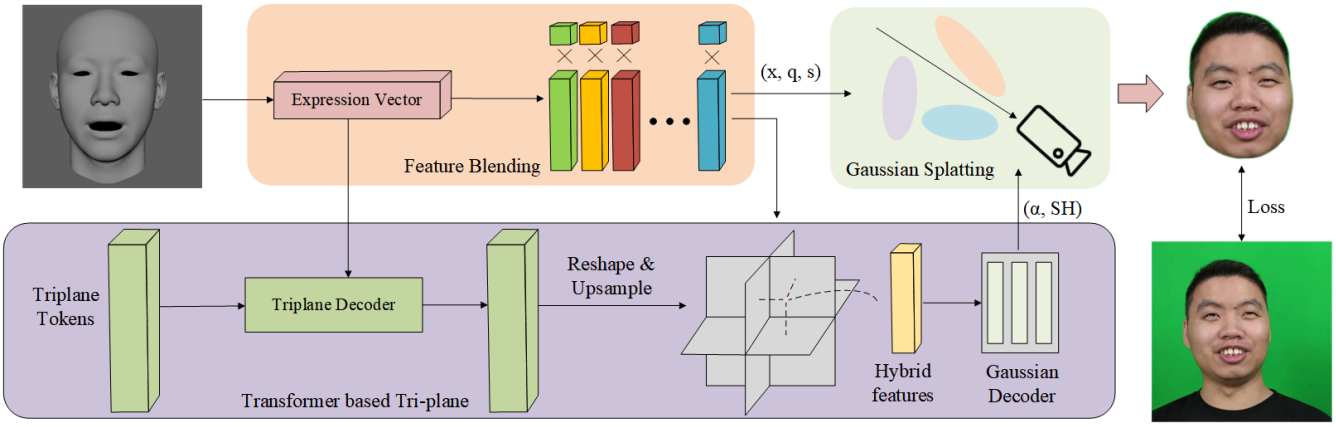

Fig. 2. Method overview. TGAvatar process begins with the random initialization of a set of Gaussians with pose, rotation, and scale bases (P, Q, S) and bias terms (p_0, q_0, s_0). In addition, a transformer based tri-plane module is employed to ensure high-fidelity novel view synthesis. Specifically, we first use a transformer-based tri-plane decoder to predict tri-plane features. Subsequently, we incorporate a tri-plane module to extract hybrid features based on the pose of each Gaussian. Finally, these hybrid features are fed into an MLP network to infer opacity (α) and spherical harmonics coefficients(SH) in each gaussian.

The primary advantage of the tri-plane representation is its efficiency. By shifting the bulk of expressive power into the explicit features and keeping the decoder small, it significantly reduces the computational cost compared to fully implicit MLP architectures without sacrificing expressiveness. This representation scales with $O(N^2)$ for feature planes, as opposed to $O(N^3)$ for dense voxel grids, allowing higher resolution features and greater detail within the same memory footprint [19].

Gaussian Splatting: 3D Gaussian Splatting (3DGS) utilizes anisotropic 3D Gaussian primitives to explicitly represent the underlying structure of a scene [1]. Each Gaussian is defined by its position (mean) $\mathbf{x}$ and a 3D covariance matrix Σ in world coordinates:

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

The covariance matrix Σ can be decomposed into a scaling matrix $\mathbf{S}$ and a rotation matrix $\mathbf{R}$, such that $\Sigma = \mathbf{RSS}^T\mathbf{R}^T$. For optimization purposes, we represent $\mathbf{S}$ with a scaling vector $\mathbf{s}$ and $\mathbf{R}$ with a unit quaternion $\mathbf{q}$. Therefore, the Gaussian function can be rewritten as $G(\mathbf{x}, \mathbf{q}, \mathbf{s})$.

To render these 3D Gaussians onto a 2D image plane, we transform the covariance matrix into camera coordinates. This is achieved using the view transformation matrix $\mathbf{W}$ and the Jacobian matrix $\mathbf{J}$, which approximates the projective transformation [43], [44]:

$$\Sigma' = \mathbf{JW}\Sigma\mathbf{W}^T\mathbf{J}^T. \tag{2}$$

The appearance of each Gaussian is influenced by two additional parameters: opacity α and spherical harmonics coefficients Y_{lm}, which, when combined with the spherical harmonics basis functions, represent view-dependent color. The color C of a pixel on the camera plane is computed by blending N ordered 3D Gaussians that overlap the pixel:

$$C = \sum_{i \in N} c_i \alpha_i \prod_{j=1}^{i-1}(1 - \alpha_j), \tag{3}$$

where c_i and α_i are the color and opacity of the i-th Gaussian, respectively.

During the training process, the Gaussians undergo alternating densification or sparsification to address under- or over-reconstruction. This adaptive density control ensures an accurate and efficient representation of the scene by dynamically adjusting the number and distribution of Gaussians.

The optimization process involves updating the position $\mathbf{x}$, the scaling vector $\mathbf{s}$, the unit quaternion $\mathbf{q}$, the opacity α, and the spherical harmonics coefficients Y_{lm} to minimize the reconstruction error. The optimization is driven by a loss function that combines L_1 and perceptual loss components to achieve high visual fidelity.

B. Feature Blending

Several existing approaches [15], [17] using MLP modules to infer parameter offsets for Gaussian components. However, these approaches often fall short in flexibly capturing the subtlety of facial expressions. These methods may inadvertently impose rigid constraints on how facial attributes are modeled, thereby limiting the expressiveness achievable in animations. Inspired by 3DMM, our approach employs a feature blending technique within each Gaussian to enhance the animation and rendering process. Our method blends pose, rotation, and scale in each Gaussian. Intuitively, opacity and spherical harmonics are relatively complex and cannot be easily mixed, so we use a tri-plane approach to predict them more precisely. Subsequent experiments also demonstrate that this approach is reasonable.

Specifilly, each 3D Gaussian in our framework contains a set of pose, rotation, and scale bases, i.e. $X \in \mathbb{R}^{N\times 3}$, $Q \in \mathbb{R}^{N\times 4}$, $S \in \mathbb{R}^{N\times 3}$, and bias terms p_0, q_0, s_0, instead of the pose, rotation and scale in original gaussian splatting. For a given frame i, we utilize the expression weights e_i to blend the pose, rotation, and scale bases P, R, S into frame-specific coefficients P_i, R_i, S_i. This process is mathematically

represented as:

$$\begin{aligned} P_i &= P^{\mathrm{T}} e_i + p_0, \\ Q_i &= Q^{\mathrm{T}} \cdot e_i + q_0, \\ S_i &= S^{\mathrm{T}} \cdot e_i + s_0 \end{aligned} \tag{4}$$

Here, $P, R,$ and S are the base pose, rotation, and scale bases, p_0, q_0, s_0 denotes the pose, rotation and scale bases, respectively, and e_i are the expression weights for frame i. In our approach, we directly add the quaternions for each Gaussian and then normalize them. To ensure accurate and meaningful results, we adjust each quaternion to ensure the inner products are positive. This adjustment is necessary because quaternions q and $-q$ represent the same rotation, and without this step, the averaging process could lead to incorrect results. In our expriments, we found that the quaternion bases are relatively close to each other in the rotation space. Thus, adding them directly and normalizing the result will not cause significant issues, making the method effective and yielding satisfactory results.

C. Transformer based Tri-plane Gaussian

Existing approaches [17], [18] treat the alpha and spherical harmonic parameters the same as other parameters, which may result in a lack of flexibility in rendering subtle facial features. GaussianHead [15] employs a tri-plane [19] to inference the alpha and spherical harmonic parameters, but it does not establish a connection with expression coefficients, and may lead to inaccurate expressions rendering. In order to address these issues, we employ a transformer-based tri-plane decoder to generate tri-plane features, which are then combined with Gaussian splatting for efficient and high-quality 3D reconstruction and rendering. This section details the steps involved in our hybrid tri-plane Gaussian approach.

The tri-plane decoder is a 6-layer transformer architecture to decode the tri-plane features using expression conditions from a fixed number of learnable tri-plane embeddings $\{f_i\}_t$, similar to other transformer architecture designs [45], each transformer block consists of a self-attention layer, a cross-attention layer, and a feed-forward layer. The expression vector captures the current frame's specific details, while the tri-plane token provides a learnable parameter set that helps in predicting the tri-plane features. The tri-plane decoder combines these inputs to predict the tri-plane features. This prediction is achieved through a series of transformer layers that utilize the cross-attention mechanism to effectively condition the tri-plane tokens on the expression vector. The output of the tri-plane decoder is subsequently reshaped and upsampled to obtain the final tri-plane representation, which consists of three orthogonal feature planes: T_{xy}, T_{xz}, and T_{yz}.

Next, the Gaussians, obtained from the feature blending process described in the previous section, are used to query the tri-plane. For each Gaussian, based on its spatial coordinates (x, y, z), we query the corresponding feature vectors from the tri-plane. This involves projecting the 3D coordinates onto the three orthogonal planes and performing trilinear interpolation to retrieve the feature vectors F_{xy}, F_{xz}, and F_{yz}. These vectors are then concatenated to form a single hybrid feature vector for each Gaussian.

Once the hybrid features are obtained, they are fed into a multi-layer perceptron (MLP) network to predict the opacity α and the spherical harmonics (SH) coefficients. The MLP takes the hybrid features as input and outputs the required parameters for rendering the Gaussians with accurate lighting and opacity effects. This step ensures that the complex properties of opacity and spherical harmonics are precisely predicted, leveraging the detailed feature representation provided by the tri-plane.

D. Training

Unlike original 3D Gaussian splatting [1], each Gaussian in our framework contains a set of pose, rotation, and scale bases, i.e., $P \in \mathbb{R}^{N\times 3}$, $Q \in \mathbb{R}^{N\times 4}$, $S \in \mathbb{R}^{N\times 3}$, and bias terms p_0, q_0, and s_0. The opacity and SH coefficients are predicted using the tri-plane module described in Section III-C, rather than being directly optimized. We train these variables in each gaussian, together with other trainable parameters in our pipeline, including the tri-plane token, the tri-plane decoder and the MLP network which used to predict opacity and SH coefficients, respectively.

Our loss function comprises a combination of L1 loss, SSIM loss, and an additional perceptual loss term, leading to the following total loss:

$$L_{\text{total}} = L_1(I_r, I_{gt}) + \lambda_s L_{\text{SSIM}}(I_r, I_{gt}) + \lambda_p L_p(I_r, I_{gt}) \tag{5}$$

Here, L_1 denotes the L1 loss, L_{SSIM} represents the Structural Similarity Index (SSIM) loss, and L_p is the perceptual loss based on a pre-trained VGG network. The weights λ_s and λ_p are hyperparameters set to balance the contributions of the SSIM and perceptual losses.

We start training from 10K uniformly random Gaussians inside the preset volume. Adaptive densification and pruning mechanisms are integrated into our training to maintain an effective representation of the scene. We follow the strategies outlined in previous work on 3D Gaussian Splatting [1], where transparent Gaussians (those with opacity α below a threshold τ_α) are pruned, and regions requiring more detail are densified by either cloning or splitting Gaussians based on positional gradients and Gaussian size.

IV. Experiments

In this section, we outline the evaluation protocol and the experiments conducted to assess the effectiveness of the proposed method. Subsequently, we present the results across three different scenarios: head reconstruction, novel view synthesis, and cross-subject expression driving. Finally, we perform an ablation study to demonstrate the contribution of each module in the proposed method.

To ensure a fair comparison, we obtained our data from public subjects as described by [29], [46]. Each subject's training dataset consisted of roughly 3000 to 4000 frames, while the test dataset comprised the final 3% to 5% of frames. Specifically, the data for each subject included four components: RGB head images with a resolution of 512×512, expression parameters from 3DMM model fitting [47], camera

TABLE I
QUANTITATIVE COMPARISON OF DIFFERENT METHODS. OUR TGAVATAR METHOD ACHIEVES THE BEST PERFORMANCE ACROSS ALL METRICS.

Method	L1↓	PSNR↑	SSIM↑	LPIPS↓
GaussianBlendshapes [18]	0.0089	32.74	0.949	0.079
INSTA [46]	0.0135	29.56	0.913	0.139
FlashAvatar [17]	0.0094	31.86	0.936	0.088
TGAvatar (Ours)	**0.0083**	**33.55**	**0.953**	**0.072**

parameters, and binary masks. The first three components were taken from open datasets, whereas the binary masks were generated using MODNet [48]. We used these binary masks in the input frames to eliminate the background. For head movements, the head was fixed in the coordinate system, and we simulated changes in head poses using camera poses [9], [29].

Our learning rates for the pose(base P and bias term p_0), rotation(base Q and bias term q_0), scale S(base S and bias term s_0), tri-plane token $\{f_i\}_t$, tri-plane decoder and the MLP network are namely 0.00016, 0.005, 0.001, 0.001, 0.0001, and 0.0001. The L_p loss, which is derived from a VGG network [49], is assigned a weight of $\lambda_p = 0.1$, while $\lambda_s = 0.2$. To avoid interference with the photometric loss during the initial training phase, the L_p loss is activated only after 10,000 iterations. Densification and pruning begins after 500 iterations and concludes at 10,000 iterations. In our experiments, we use a Spherical Harmonics (SH) degree of $k = 3$. Our models are trained on a single NVIDIA GTX3090 GPU for a total of 100,000 iterations, which takes approximately 2 hours. In testing, our method achieves an inference speed of approximately 50 fps.

A. Head Reconstruction

The image in Figure 3 compares the visual performance of different methods. INSTA [46], FlashAvatar [17] and GaussianBlendshapes [18], and our TGAvatar, against the ground truth (GT) in rendering facial avatars under various expressions and details. The performance of INSTA is inadequate, especially for expressions involving squinting and pouting, where it fails to accurately capture the facial deformations, leading to noticeable artifacts and unrealistic renderings. Additionally, INSTA struggles to render fine details such as teeth, and has an obviously smoothing effect in facial details. FlashAvatar, while generally better, also shows some artifacts, particularly on teeth and details around the mouth. As evidence, please refer to the head reconstructed by FlashAvatar in row 1 and row 4. Similar to FlashAvatar, GaussianBlendshapes also failed to accurately repersent the facial details such as teeth and eyeglasses, and has an ghosting effect around teeth(row 1) and an obvious artifact on eyeglasses(row 3). In constract, our TGAvatar outperforms the above methods, accurately captures fine details such as teeth, eye squint, reflections and slight details around the face, providing a more lifelike and accurate depiction, closely matching the ground truth. This level of detail and accuracy is challenging for previous head avatar approaches trained on monocular videos.

The quantitative results shown in Table I further corroborate the visual superiority of our TGAvatar method. Our approach achieves the best performance across multiple metrics, with the lowest L1 error (0.0083), the highest PSNR (33.55), the highest SSIM (0.953), and the lowest LPIPS (0.072). These results indicate that our method produces avatars with the highest fidelity, structural similarity, and perceptual quality compared to other methods.

B. Novel View Synthesis

Figure 4 demonstrates the model's capability to maintain visual coherence and detail across different views. Each novel view generated by the TGAvatar retains the structural integrity and detailed texture of the facial features, such as the skin pores and subtle facial expressions, which are often lost in the outputs from other models.

C. Cross-Subject Expression Driving

In the cross-subject expression driving experiment, we assess the ability of various methods to transfer facial expressions from a source subject to a target subject while preserving the target's unique identity. Figure 5 compares the results of four different methods: INSTA [46], FlashAvatar [17] and GaussianBlendshapes [18], and our proposed TGAvatar. The first column shows the ground truth (GT) expressions, followed by the results of each method.

The expressions range from pouting and smiling to wide-open mouth and other complex facial movements. The TGAvatar method demonstrates superior performance in maintaining the fine details and realistic appearance of the target subject's expressions, outperforming other methods that exhibit artifacts, smoothing effects, and loss of facial detail. For example, in the first row, our method captures the subtle details of a pouting expression and eye blinking, which are often challenging to replicate, while other methods show lack of mouth details. In the fourth row, our method also performs better on teeth and facial details rendering. This experiment highlights the effectiveness of TGAvatar in producing high-fidelity and lifelike facial animations.

TABLE II
ABLATION STUDY RESULTS. REMOVING ANY KEY COMPONENT LEADS TO A SIGNIFICANT DROP IN PERFORMANCE, HIGHLIGHTING THEIR IMPORTANCE IN OUR FRAMEWORK.

Meth	L1↓	PSNR↑	SSIM↑	LPIPS↓
Ours w/o tri-plane	0.0095	31.15	0.911	0.082
Ours w/o feature blending	0.0115	30.72	0.892	0.095
Ours w/o L_p	0.0086	32.95	0.934	0.079
Ours	**0.0083**	**33.55**	**0.953**	**0.072**

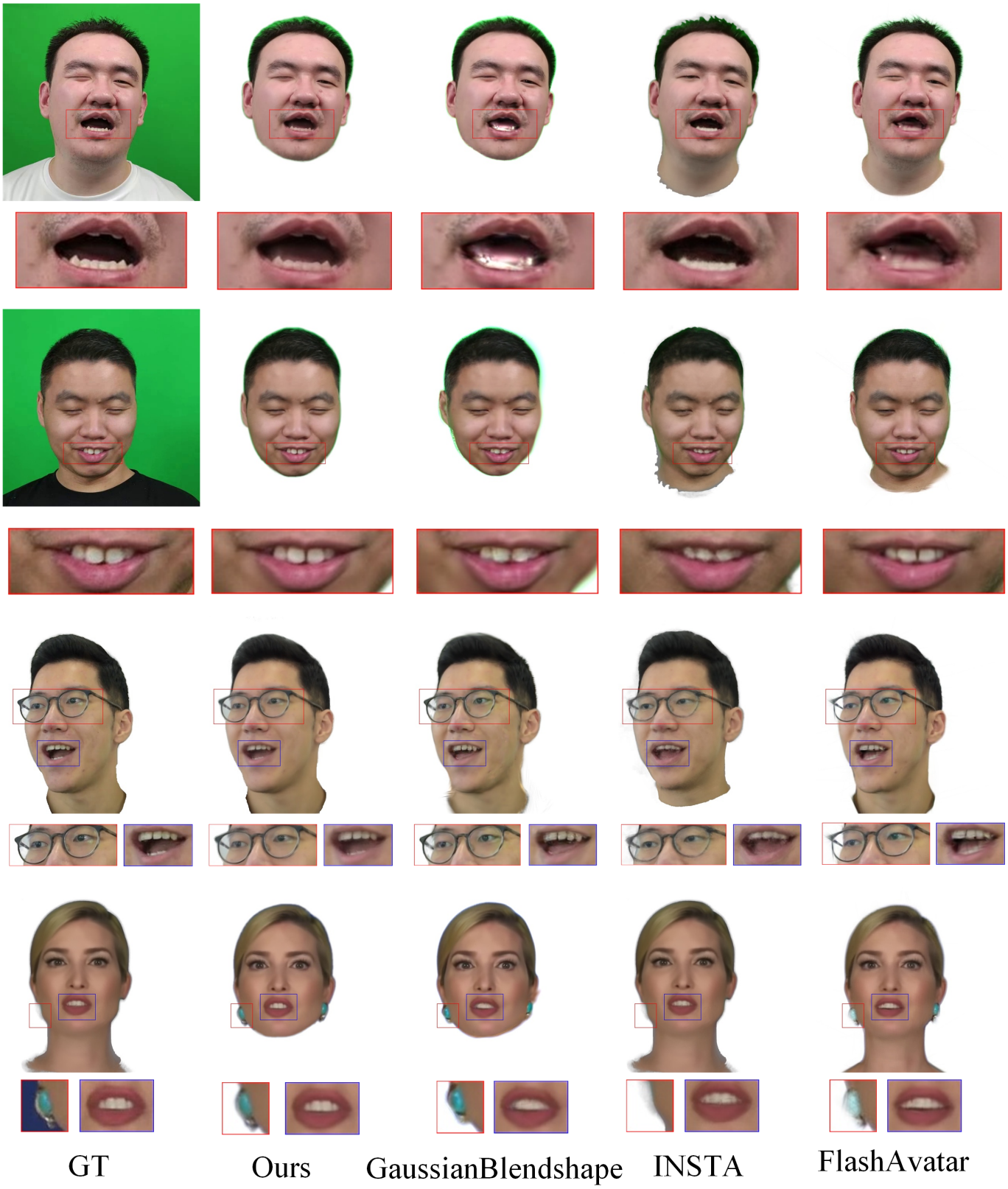

Fig. 3. Qualitative comparisons between our TGAvatar and INSTA [46], FlashAvatar [17] and GaussianBlendshapes [18]. Results are executed under the configurations specified in their works. For INSTA dataset, INSTA and GaussianBlendshapes provide pretrained models, therefore, these results are evaluated by their pretrained models. Our TGAvatar achieves better results, particularly in capturing details such as teeth, eyes, wrinkles and reflections.

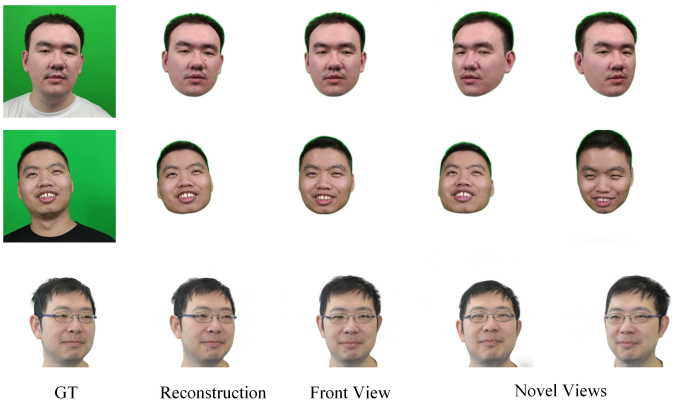

Fig. 4. In qualitative results on novel view synthesis, TGAvatar generates consistent rendering results across these novel views.

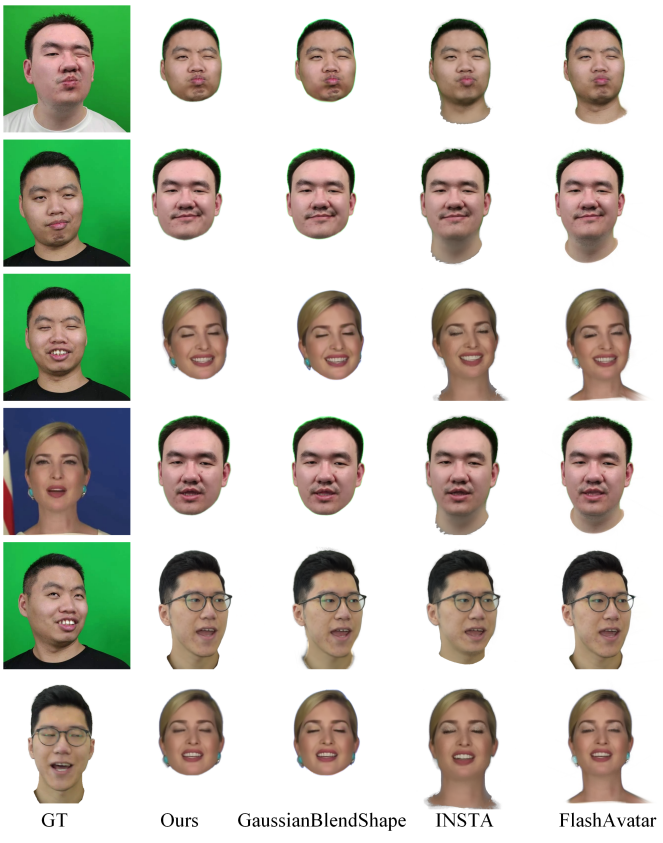

Fig. 5. Cross-subject expression driving experiment, various methods are compared in their ability to transfer expressions from a source subject to a target subject while maintaining the target's unique identity. The qualitative results show that our proposed TGAvatar method outperforms INSTA [46], FlashAvatar [17] and GaussianBlendshapes [18], accurately capturing fine facial details and providing lifelike, realistic results with minimal artifacts.

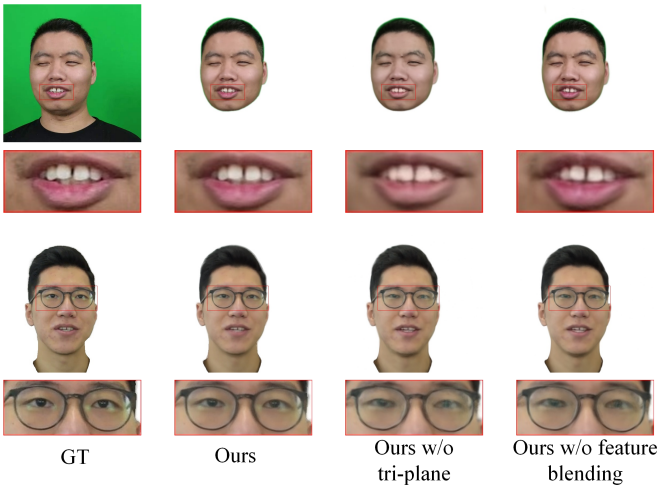

Fig. 6. Qualitative results of the ablation study. The first row focuses on facial expressions involving lip movements, and the second row highlights the rendering of teeth and smile details. The full model outperforms both ablated versions, accurately capturing fine details and complex textures.

D. Ablation Study

To further validate the effectiveness of the key components in our TGAvatar framework, we conducted an ablation study. This study focuses on assessing the impact of three primary components: the tri-plane module, the feature blending technique, and the perceptual loss. We systematically removed each component and evaluated the resulting performance.

Firstly, we removed the tri-plane module responsible for inferring opacity and SH coefficients and instead used the feature blending technique to estimate these parameters directly. We refer to this model as Ours w/o tri-plane. Secondly, we replaced the feature blending technique with an MLP, which takes the expression coefficients and the pose, rotation, and scale of each gaussian as input, and predicts the delta pose, rotation, and scale values. We refer to this model as Ours w/o feature blending. Finally, we assessed the effect of excluding the perceptual loss from our training objective. We refer to this model as Ours w/o L_p.

The qualitative results of our ablation study are depicted in Figure 6, which illustrates the differences in performance between the full model and the ablated versions. For example, in the second row, the comparison focuses on the rendering of teeth and smile details. The full model accurately captures the fine details of the teeth and the subtle nuances of the smile. However, in the *Ours w/o tri-plane* model, the teeth appear less defined, and the overall texture is more smoothed out, showing the importance of the tri-plane module in maintaining detail and texture quality. The *Ours w/o feature blending* model exhibits even greater blurring and a loss of structural integrity in the teeth, further highlighting the necessity of feature blending for high-quality avatar reconstruction.

The quantitative results of our ablation study are presented in Table II. Removing any key component led to a significant drop in performance, highlighting their importance in our framework. Without the tri-plane module, the method struggled to capture fine details and complex lighting effects

accurately. Replacing the feature blending technique with an MLP resulted in even more significant performance degradation. Excluding the perceptual loss also led to noticeable declines in performance.

V. Limitations and Ethical Considerations

While our TGAvatar framework presents significant advancements in 3D head animation and reconstruction, it is not without limitations. One notable limitation is the framework's dependency on the diversity of the training data, particularly in terms of head poses. If certain views, such as side profiles, are underrepresented or entirely absent in the training data, the performance of our model may degrade significantly when generating or reconstructing these unseen angles. This can result in less accurate or distorted renderings, particularly in areas with complex geometry or occlusions, such as the ears or the side contours of the face.

From an ethical standpoint, the ability to create highly realistic 3D avatars raises important considerations regarding privacy and consent. The misuse of such technology could lead to scenarios where individuals' likenesses are replicated without their permission, potentially leading to identity theft or other forms of digital manipulation. It is crucial that researchers and developers in this field adhere to strict ethical guidelines, ensuring that the creation and use of digital avatars are conducted with explicit consent and with respect to individuals' privacy rights. Furthermore, the potential for deepfakes and other forms of synthetic media generated by such technology necessitates ongoing discussions and regulations to prevent misuse.

VI. Conclusion

In this paper, we introduced TGAvatar, a novel framework that leverages the power of 3D Gaussian Splatting and advanced modeling techniques to achieve high-fidelity 3D head animation and reconstruction. By integrating a tri-plane module for precise opacity and spherical harmonics inference and employing a feature blending technique inspired by 3D Morphable Models, our method significantly enhances the realism and detail of facial avatars, while rendering on real-time frame rates. Our comprehensive evaluations demonstrate that TGAvatar not only surpasses existing methods in terms of rendering quality and computational efficiency but also sets a new benchmark for the field. Our future work will focus on further optimizing the method for broader applicability, enhancing robustness under diverse conditions.

References

[1] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–14, 2023.

[2] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194.

[3] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *2009 sixth IEEE international conference on advanced video and signal based surveillance*. Ieee, 2009, pp. 296–301.

[4] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans." *ACM Trans. Graph.*, vol. 36, no. 6, pp. 194–1, 2017.

[5] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[6] K. Zhang, G. Riegler, N. Snavely, and V. Koltun, "Nerf++: Analyzing and improving neural radiance fields," *arXiv preprint arXiv:2010.07492*, 2020.

[7] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5855–5864.

[8] S. J. Garbin, M. Kowalski, M. Johnson, J. Shotton, and J. Valentin, "Fastnerf: High-fidelity neural rendering at 200fps," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 14 346–14 355.

[9] G. Gafni, J. Thies, M. Zollhofer, and M. Nießner, "Dynamic neural radiance fields for monocular 4d facial avatar reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8649–8658.

[10] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla, "Nerfies: Deformable neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5865–5874.

[11] J. Luiten, G. Kopanas, B. Leibe, and D. Ramanan, "Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis," *arXiv preprint arXiv:2308.09713*, 2023.

[12] Z. Yang, X. Gao, W. Zhou, S. Jiao, Y. Zhang, and X. Jin, "Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction," *arXiv preprint arXiv:2309.13101*, 2023.

[13] G. Wu, T. Yi, J. Fang, L. Xie, X. Zhang, W. Wei, W. Liu, Q. Tian, and X. Wang, "4d gaussian splatting for real-time dynamic scene rendering," *arXiv preprint arXiv:2310.08528*, 2023.

[14] Z. Yang, H. Yang, Z. Pan, X. Zhu, and L. Zhang, "Real-time photo-realistic dynamic scene representation and rendering with 4d gaussian splatting," *arXiv preprint arXiv:2310.10642*, 2023.

[15] J. Wang, J.-C. Xie, X. Li, F. Xu, C.-M. Pun, and H. Gao, "Gaussianhead: High-fidelity head avatars with learnable gaussian derivation," 2024.

[16] Y. Chen, L. Wang, Q. Li, H. Xiao, S. Zhang, H. Yao, and Y. Liu, "Monogaussianavatar: Monocular gaussian point-based head avatar," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–9.

[17] J. Xiang, X. Gao, Y. Guo, and J. Zhang, "Flashavatar: High-fidelity head avatar with efficient gaussian embedding," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[18] S. Ma, Y. Weng, T. Shao, and K. Zhou, "3d gaussian blendshapes for head avatar animation," in *ACM SIGGRAPH Conference Proceedings, Denver, CO, United States, July 28 - August 1, 2024*, 2024.

[19] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis *et al.*, "Efficient geometry-aware 3d generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 123–16 133.

[20] J. Sun, X. Wang, L. Wang, X. Li, Y. Zhang, H. Zhang, and Y. Liu, "Next3d: Generative neural texture rasterization for 3d-aware head avatars," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 20 991–21 002.

[21] S. An, H. Xu, Y. Shi, G. Song, U. Y. Ogras, and L. Luo, "Panohead: Geometry-aware 3d full-head synthesis in 360deg," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 950–20 959.

[22] J. Booth, A. Roussos, E. Ververas, E. Antonakos, S. Ploumpis, Y. Panagakis, and S. Zafeiriou, "3d reconstruction of "in-the-wild" faces in images and videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 11, pp. 2638–2652, 2018.

[23] B. Egger, W. A. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani *et al.*, "3d morphable face models—past, present, and future," *ACM Transactions on Graphics (ToG)*, vol. 39, no. 5, pp. 1–38, 2020.

[24] A. Tewari, M. Zollhofer, H. Kim, P. Garrido, F. Bernard, P. Perez, and C. Theobalt, "Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction," in *Proceedings of the IEEE international conference on computer vision workshops*, 2017, pp. 1274–1283.

[25] L. Tran and X. Liu, "Nonlinear 3d face morphable model," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7346–7355.
[26] X. Tu, J. Zhao, M. Xie, Z. Jiang, A. Balamurugan, Y. Luo, Y. Zhao, L. He, Z. Ma, and J. Feng, "3d face reconstruction from a single image assisted by 2d face images in the wild," *IEEE Transactions on Multimedia*, vol. 23, pp. 1160–1172, 2020.
[27] Y. Li, Q. Hao, J. Hu, X. Pan, Z. Li, and Z. Cui, "3d3m: 3d modulated morphable model for monocular face reconstruction," *IEEE Transactions on Multimedia*, vol. 25, pp. 6642–6652, 2022.
[28] S. Athar, Z. Shu, and D. Samaras, "Flame-in-nerf: Neural control of radiance fields for free view face animation," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2023, pp. 1–8.
[29] X. Gao, C. Zhong, J. Xiang, Y. Hong, Y. Guo, and J. Zhang, "Reconstructing personalized semantic facial nerf models from monocular video," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 6, pp. 1–12, 2022.
[30] S. Athar, Z. Xu, K. Sunkavalli, E. Shechtman, and Z. Shu, "Rignerf: Fully controllable neural 3d portraits," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 364–20 373.
[31] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. Zhang, "Ad-nerf: Audio driven neural radiance fields for talking head synthesis," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5784–5794.
[32] Z. Bai, F. Tan, Z. Huang, K. Sarkar, D. Tang, D. Qiu, A. Meka, R. Du, M. Dou, S. Orts-Escolano *et al.*, "Learning personalized high quality volumetric head avatars from monocular rgb videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 890–16 900.
[33] H.-B. Duan, M. Wang, J.-C. Shi, X.-C. Chen, and Y.-P. Cao, "Bakedavatar: Baking neural fields for real-time head avatar synthesis," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 6, pp. 1–17, 2023.
[34] X. Liu, Y. Xu, Q. Wu, H. Zhou, W. Wu, and B. Zhou, "Semantic-aware implicit neural audio-driven video portrait generation," in *European conference on computer vision*. Springer, 2022, pp. 106–125.
[35] Y. Xu, L. Wang, X. Zhao, H. Zhang, and Y. Liu, "Avatarmav: Fast 3d head avatar reconstruction using motion-aware neural voxels," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–10.
[36] Y. Xu, H. Zhang, L. Wang, X. Zhao, H. Huang, G. Qi, and Y. Liu, "Latentavatar: Learning latent expression code for expressive neural head avatar," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–10.
[37] S. Shen, W. Li, X. Huang, Z. Zhu, J. Zhou, and J. Lu, "Sd-nerf: Towards lifelike talking head animation via spatially-adaptive dual-driven nerfs," *IEEE Transactions on Multimedia*, 2023.
[38] A. Cao and J. Johnson, "Hexplane: A fast representation for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 130–141.
[39] S. Fridovich-Keil, G. Meanti, F. R. Warburg, B. Recht, and A. Kanazawa, "K-planes: Explicit radiance fields in space, time, and appearance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 479–12 488.
[40] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM transactions on graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.
[41] Z. Ma, X. Zhu, G.-J. Qi, Z. Lei, and L. Zhang, "Otavatar: One-shot talking face avatar with controllable tri-plane rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 901–16 910.
[42] X. Zhao, L. Wang, J. Sun, H. Zhang, J. Suo, and Y. Liu, "Havatar: High-fidelity head avatar via facial model conditioned neural radiance field," *ACM Transactions on Graphics*, vol. 43, no. 1, pp. 1–16, 2023.
[43] M. Zwicker, H. Pfister, J. Van Baar, and M. Gross, "Surface splatting," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001, pp. 371–378.
[44] Zwicker, Matthias and Pfister, Hanspeter and Van Baar, Jeroen and Gross, Markus, "Ewa splatting," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 3, pp. 223–238, 2002.
[45] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General perception with iterative attention," in *International conference on machine learning*. PMLR, 2021, pp. 4651–4664.
[46] W. Zielonka, T. Bolkart, and J. Thies, "Instant volumetric head avatars," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4574–4584.
[47] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Luthi, S. Schönborn, and T. Vetter, "Morphable face models-an open framework," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 75–82.
[48] Z. Ke, J. Sun, K. Li, Q. Yan, and R. W. Lau, "Modnet: Real-time trimap-free portrait matting via objective decomposition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 1140–1147.
[49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

Ruigang Hu is currently a master candidate in the College of Electronics and Information Engineering, Tongji University. His research interests include computer vision, machine learning and neural rendering(e.g. NeRF, 3DGS).

Xuekuan Wang received the B.S. and M.S. degrees from the College of Electronics and Information Engineering, Tongji University, in 2014 and 2017, respectively. Currently, he is a senior algorithm engineer at Baidu. His research interests include computer vision, pattern recognition and machine learning, in particular, focusing on person re-identification for visual surveillance.

Yichao Yan received the BE and PhD degrees in electrical engineering from Shanghai Jiao Tong University, in 2013 and 2019, respectively. He is currently an assistant professor with the AI Institute, Shanghai Jiao Tong University. He has authored/coauthored more than 20 peer-reviewed papers, including those in highly regarded journals and conferences such as IEEE Transactions on Pattern Analysis and Machine Intelligence, CVPR, ECCV, ACMMM, IJCAI, TMM, etc. His research interests include 3D generation, video analysis, and deep learning.

Cairong Zhao is currently a Professor of the College of Electronic and Information Engineering at Tongji University. He received a Ph.D. degree from Nanjing University of Science and Technology, an M.S. degree from Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences and a B.S. degree from Jilin University, in 2011, 2006 and 2003, respectively. He works on visual and intelligent learning, including computer vision, pattern recognition and visual surveillance. He has published over 50 top-rank international conferences and journals in the field, including CVPR, ICCV, ECCV, ICML, NIPS, ICLR, AAAI, ACM MM, TPAMI, IJCV, TIP, and TIFS, etc.. He serves as the Director of the Computer Vision Specialized Committee of the Shanghai Computer Society, Deputy Secretary-General of the Young Professionals Committee of the China Society of Image and Graphics, and Deputy Secretary-General of the Pattern Recognition and Machine Intelligence Specialized Committee of the China Automation Society, etc..