This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

VoluMe – Authentic 3D Video Calls from Live Gaussian Splat Prediction

Martin de La Gorce Charlie Hewitt Tibor Takács Robert Gerdisch Zafiirah Hosenie
Givi Meishvili Marek Kowalski Thomas J. Cashman Antonio Criminisi

Microsoft, Cambridge, UK

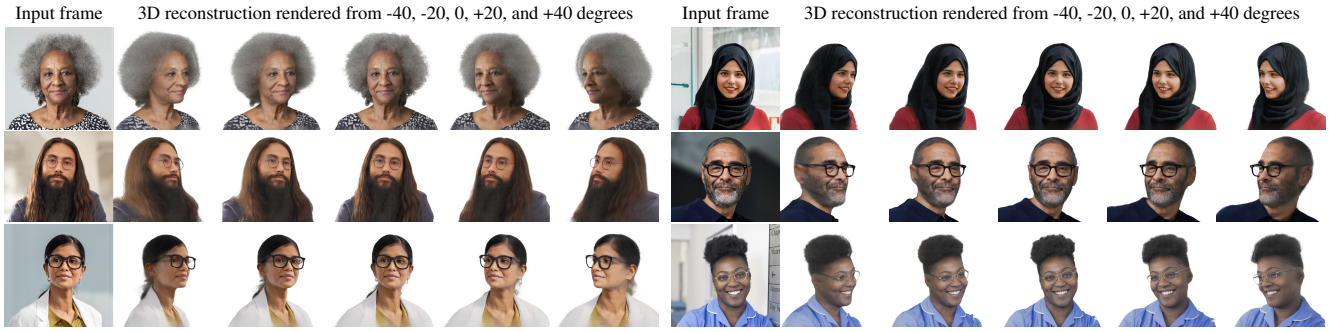


Figure 1. Given a streaming monocular video of a person, our method reconstructs each frame as a realistic 3D Gaussian splat of the moment captured by that frame. Avatar systems build a model of the person's appearance in the past, and may not fully adapt to how a person looks in the present. Our method builds a realistic representation in real time and without enrolment. We use the term *authenticity* for the property that the input frame is faithfully represented, shown in this figure by the match between the input frame and the reconstruction when rendered to the original camera's viewpoint. Our method faithfully represents eye glasses, headwear and diverse head and facial hair.

Abstract

Virtual 3D meetings offer the potential to enhance copresence, increase engagement and thus improve effectiveness of remote meetings compared to standard 2D video calls. However, representing people in 3D meetings remains a challenge; existing solutions achieve high quality by using complex hardware, making use of fixed appearance via enrolment, or by inverting a pre-trained generative model. These approaches lead to constraints that are unwelcome and ill-fitting for videoconferencing applications.

We present the first method to predict 3D Gaussian reconstructions in real time from a single 2D webcam feed, where the 3D representation is not only live and realistic, but also authentic *to the input video. By conditioning the 3D representation on each video frame independently, our reconstruction faithfully recreates the input video from the captured viewpoint (a property we call* authenticity*), while generalizing realistically to novel viewpoints. Additionally, we introduce a stability loss to obtain reconstructions that are temporally stable on video sequences.*

We show that our method delivers state-of-the-art accuracy in visual quality and stability metrics compared to existing methods, and demonstrate our approach in live one-to-one 3D meetings using only a standard 2D camera and display. This demonstrates that our approach can allow anyone to communicate volumetrically, via a method for 3D videoconferencing that is not only highly accessible, but also realistic and authentic.

1. Introduction

Online video meetings have transformed the ability for people to communicate and collaborate while physically remote, but standard 2D video meetings create challenges with fatigue [24], make it difficult to infer where attention is directed [28], and have limited effectiveness [22]. Recent evidence suggests that virtual 3D meetings offer the potential to enhance copresence, increase engagement, and thus improve communication effectiveness [13, 28, 34]. However, representing people in 3D meetings, with accuracy and realism in both appearance and movement, and on consumer devices, is an unsolved challenge.

Existing solutions achieve high quality either with complicated hardware, such as multiple input cameras, dedicated sensors and powerful processors, e.g., Project Starline [13]. Alternatively, photoreal avatars [26] use a latent represen-

tation to animate a pre-enrolled model of the user, limiting the appearance to what was seen at enrolment time, and thus intentionally failing to adapt to the current moment. Finally, for approaches such as Live 3D Portrait [35] and TriPlaneNet [1], for which the 3D reconstruction is learned or predicted from the EG3D generative model [4], the 3D reconstructions are limited by the textures and geometry learned by that model (see Fig. 9). In each of these approaches, the constraints on the 3D representations are unwelcome for videoconferencing, where users need the rich, authentic and unconstrained representation that is familiar from 2D video, without the need for specialized hardware. In order to meet this challenge, we need a 3D representation that is

1. **Authentic:** generates images that match the input video foreground (with every detail of, for example, clothes, hair, and glasses from the present moment), when rendered from the original camera's viewpoint;
2. **Realistic:** generates plausible and realistic reconstructions when rendered in new views, while supporting the full diversity of human communication (e.g. including diverse hair, headwear and accessories);
3. **Live:** runs in real time on consumer devices;
4. **Stable:** generates predictions that are stable with respect to time and as the viewpoint changes, to avoid flickering.

. In this paper, we demonstrate the first method to satisfy all four properties. Examples of our results are shown in Fig. 1. Note that we can do so *without* the need for a full 360° reconstruction, while addressing an important class of 3D video meetings (the capture volume for Project Starline [13], for example, necessitates novel views only up to 55° from frontal). In this paper, our contributions are:

- We demonstrate 3D Gaussian reconstructions from 2D video that are live, authentic and realistic for a range of novel views that are up to 40° offset from the input.
- We show that these reconstructions are competitive with the state of the art for realism and quality, and excel at reconstructing the input view, even for a wide range of unconstrained inputs typical of videoconferencing scenarios (such as diverse hair, lighting, eyewear, clothing and accessories). We demonstrate state-of-the-art results on PSNR and LPIPS metrics on the Ava256 and Cafca datasets, and better results on jitter metrics on the Ava256 dataset.
- We demonstrate a real-time system that uses these 3D reconstructions to enable 3D videoconferencing at 30FPS with simulated motion parallax, using only a standard RGB camera and screen.

1.1. Related work

Photorealistic avatars. Avatar-based techniques rely on an enrolment process, where an appearance and geometry model is built for a given user. Enrolment may involve capturing multiple views of the user from a phone [3, 12] or multi-camera rig [26]; this process may be lengthy, as the user needs to show different head positions, facial expressions etc., and may be inaccessible or expensive in the case of multi-camera setups. Recent work on photorealistic avatars relies on radiance field representations, first by adapting NeRF [21] to represent a head model using deformation fields [43] or deformable volumes [2, 6, 18]. With Gaussian splatting, Kerbl et al. [10] demonstrate that radiance fields can be rendered more efficiently via rasterization, and inspire a large range of work that adapts Gaussian splatting to digital humans [5, 7, 23, 27, 37, 39, 40].

In this work, we draw inspiration from the *realism* demonstrated by Gaussian avatars, but apply this representation to a setting where *authenticity* is important; we cannot rely on a fixed enrolled appearance from the past. Instead we want to ensure that our 3D reconstruction adapts instantly to any appearance changes, as is true for all 2D video calls today.

Single view reconstruction. A second class of related work predicts a 3D reconstruction from a single 2D image or video frame, without prior enrolment. These methods can offer high *authenticity* and *realism*, but the challenge of reconstructing 3D from a single 2D input makes it harder for these methods to be *live* and *stable*.

Live 3D Portrait [35] and TriPlaneNet [1] both make use of the EG3D generative model [4] and predict an implicit 'triplane' radiance field representation. Live 3D Portrait runs in real time and achieves impressive results, satisfying many of our requirements, but by distilling knowledge from a generative 3D model, the resulting reconstructions are constrained by said model (see Fig. 9), and TriPlaneNet predicts a latent code for EG3D, making this a hard constraint.

The most promising explicit representation methods are those that predict Gaussian splat [10] outputs, as they can capture thin structures, specularities and transparencies. Xu et al. [38] and Zou et al. [44] combine the benefits of (implicit) triplane and (explicit) Gaussian splat representations, but both approaches are based on transformer architectures that do not scale well to *live* high-resolution Gaussian splat prediction. Lyu et al. [19] train a diffusion model to predict multiple views from a single image, which are then used to predict 3D Gaussians using GS-LRM [41]; this grants high *realism*, but the multi-view diffusion prediction makes it harder to produce *stable* outputs for video sequences.

Szymanowicz et al. [32] circumvent these problems using a 2D convolutional U-Net architecture [30] that directly predicts a Gaussian splat representation, or 'splatter image', in a method that is capable of *live* reconstruction but is not applied to human reconstruction, nor to video sequences.

In this paper, we build upon the Splatter Image [32] method and demonstrate that direct Gaussian splat prediction can satisfy the requirements for *realistic*, *authentic*, *live* and *stable* 3D video calls. In contrast to Szymanowicz et al. [32], we use a planar homography to reduce perspective distortion when a person appears near the border of the input image,

we incorporate a scaling step during model training to deal with varying distances between the camera and the subject, we output two Gaussians per pixel to increase the quality of the reconstruction, and we add a stability loss to reduce temporal flicker in reconstructions of video sequences. A concurrent work [31] also uses two Gaussians per pixel, but for scene reconstruction rather than human reconstruction.

3D reconstruction with synthetic data. One interesting aspect of our work is the choice to train on purely synthetic data. This enables multi-view data with perfect camera parameters that emulates in-the-wild scenarios; all important factors for our method. However, recent work reconstructing 3D humans using methods trained on synthetic data have encountered issues generalizing to real images at test time. For example, Rodin [36] trains a generative model to create 3D avatar representations. Due to the synthetic training data their 3D results look noticeably synthetic, matching the domain of the training data. Similar issues are visible for Live 3D Portrait [35] and TriPlaneNet [1] (e.g., eyes following the camera), although the EG3D-based synthetic data has a smaller domain gap to real images. Other recent methods aim to explicitly handle the domain gap by training a generative prior model on synthetic data and fine-tuning on real enrolment data [2, 27]. This does not suit our application, as the results are neither *live* nor *authentic*. Instead we choose to follow Splatter Image [32], in which the output is much more directly tied to the input with a feed-forward network. This allows our network to generalize well to real images, despite being trained only on synthetic data. The skip connections and direct-sampling elements of the architecture are critical in allowing information to be forwarded directly from the input image to the output 3D representation, while the bottleneck can learn a strong prior over 3D geometry based on the synthetic training data.

2. Training the 3D Reconstruction Model

2.1. System Design

Our method takes a single image as input, performs face detection on that image, extracts a region around the face and runs our model on this region. The output of the model is a set of 3D Gaussians that represent the 3D shape and appearance of the face in the image. The model architecture is based on Splatter Image [32]. We also use a U-Net architecture [25] to predict the Gaussians, though we choose a smaller backbone than SongUNet [30] to reduce execution time. We further reduce the number of convolution layers to only two per resolution and increase the number of resolutions to five to maintain a large receptive field, also omitting the self-attention layers. To retain high-quality results we add four new ingredients in our method, described below.

Optimizable layers. The U-Net architecture uses only 2D convolutions and non-linear activation, which is limited in its ability to create output with high spatial frequencies. This limitation becomes stronger as we decrease the size of the network when aiming for a network that runs in real time. To mitigate this problem, we add explicit channels to the network that can take a different value for each pixel and whose values are optimized alongside the other parameters of the network. The optimizable channels enable a lightweight shallow network to generate high frequency patterns with a minor increase in compute, and so improve the quality of the final reconstruction. We add four channels at each stage of the decoding part of the U-Net architecture.



Figure 2. Splatter image components: a) input ROI, b) colour, c) opacity, d) log scale, e) z-depth, f) xyz, g) rotation (Euler angles)

Direct colour sampling. Following Szymanowicz et al. [32], we use offsets that allow 3D Gaussians to move away from the rays associated to their pixel in the splatter image. For a Gaussian that is visible with large offset, the information provided by the skip connection in the U-Net is not well collocated with the location that the Gaussian would reproject in the input image. As a result, the information most useful for predicting the Gaussian colour is not provided by the skip connection and needs to be propagated through the deeper layers, which limits the ability to maintain texture details. In order to mitigate this, we compute the 2D location of the Gaussian after applying offset and projection into the input image and then use direct bilinear sampling of the input image where the Gaussian reprojects. We provide this extra information as additional channels to a shallow convolutional residual block that predicts the Gaussian colour, scales, and rotation.

Multiple Gaussians per pixel. In order to provide more flexibility to the representation, we generate multiple Gaussians per pixel [31]. This allows the network to predict multiple layers of the surface while only increasing the computational cost of the network by a small amount. We determined experimentally that using two Gaussians per pixel provided a good quality improvement, while three did not yield noticeable improvements and significantly increased rendering time. As visible in Fig. 2 the Gaussians in the first layer have larger scales and seem to contribute more to occluded regions, while the ones in the second layer have much smaller scales and tend to represent high-frequency details in the visible regions.

ROI homography mapping. Our method runs on a region of interest (ROI) around the face, rather than the whole

image. The simplest approach to extract an ROI is a simple image crop. However, this has important limitations when it comes to predicting 3D geometry; faces away from the image centre appear distorted in the cropped images due to the perspective projection. CLIFF [16] addresses this by providing the focal length and central point of a new virtual camera with oblique frustum that corresponds to the cropped region as input to the DNN. We instead create a virtual camera with a symmetric frustum, that is rotated to point to the centre of the face in the original image, with the FOV computed such that the face covers a third of the FOV angle. We then warp the input image to match this new camera using a homography. Because this camera has a symmetric frustum, we only need to provide the normalized focal length f/w as input to the neural network. We concatenate the input image with the ray directions and two channels that contain the repeated normalized focal length and its inverse.

2.2. Data generation

Our method takes one image as input and generates a set of Gaussians that are rendered in new views using a differentiable renderer. These renders are compared to ground truth views to obtain a training loss. For training we therefore need a multi-view dataset, but capturing a diverse, multi-view, in-the-wild dataset at a scale suitable to train a Gaussian splat prediction model is a formidable challenge. Instead, we use *exclusively* synthetic data to train our method, employing the data generation pipeline described by Hewitt et al. [8]. This enables us to create multi-view data with perfect camera parameters, including diverse subjects, and emulating in-the-wild scenarios. Our synthetic dataset includes $N_s = 20{,}000$ subjects with varied face shape, skin texture, hair, clothing, glasses, body pose and facial expression.

In order for the prediction to align well with the ground truth in the novel views we need the model to predict accurate absolute positions of the Gaussians. Due to the depth ambiguity that is inherent to single-view reconstruction, this is a very challenging task. To resolve this, previous Gaussian splat prediction methods train models using synthetic data with 3D objects positioned at fixed distance to the camera [32]. This approach has critical limitations: the same face looks very different when captured from cameras at different distances; by training the model with a fixed distance, it cannot learn strong priors about the 3D proportions of faces. Further, absolute depth prediction is not required for our video call application, where small offsets are acceptable as long as the face shape is predicted correctly. Requiring the trained model to predict very accurate distances during training is therefore wasteful in terms of model capacity.

We instead resolve the depth ambiguity by rescaling the prediction during training (see Sec. 2.3). This is enabled by high-quality ground-truth depth generation from the synthetic pipeline. As such, we sample input cameras at various distances to the subject (ranging from 40cm to 1m). The input camera should emulate a webcam, so we position faces randomly within the image, ensuring that the angle of the face to the camera is within 30°, and render at 1080×720 resolution. The images used for the supervision do not need to reflect the variability in the input data. We therefore render 512×512 pixel images from ten cameras closer to the face, pointing directly at the centre of the face, and whose positions are distributed randomly on a 45° spherical cap centred around the input camera direction. Samples of our synthetic training data are shown in Fig. 3.



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

Figure 3. Example images from our synthetic training data. a) Input image with background. b-d) Three renders (out of 10) from which the foreground image and alpha mask are used for supervision. Note the presence of glasses and different hair styles.

2.3. Training loop

The set of operations done on a single multi-view image sample in a batch during the training can be summarized as follows: (1) Using an ROI around the face based on the ground-truth 2D landmarks, define the rotated camera and warp the image using the homography as described in Sec. 2.1. Provide the warped image and ROI camera parameters as input to the DNN. (2) Predict the splatter image in the ROI camera coordinate system. (3) Render the splatter image Gaussian depth map into the input view using alpha composition to get differentiable rendering and gradients. Compute a global scale correction transform centred on the camera optical centre applied to the Gaussians by comparing the depth in the overlapping region of the ground truth mask and the rendered depth mask. The scale factor is obtained through linear least squares. (4) Apply the scale and transform the Gaussians back to the original camera coordinate system, render the corrected Gaussians in the virtual views and compare to ground truth images. The process is visualized in the top of Fig. 4.

During training we minimize the following loss function

$$L = L_d + \lambda_\sigma L_\sigma + \lambda_m L_m + \lambda_c L_c + \lambda_j L_j \qquad (1)$$

with data term L_d, regularizers L_σ, L_m and L_c and stability term L_j and corresponding weights defined below.

Following Szymanowicz et al. [32] we define L_d as:

$$L_d = \frac{1}{N_v N_s} \sum_{v=1}^{N_v} \sum_{s=1}^{N_s} L_i(R(U(I_1^s), C_v), I_v^s) \qquad (2)$$

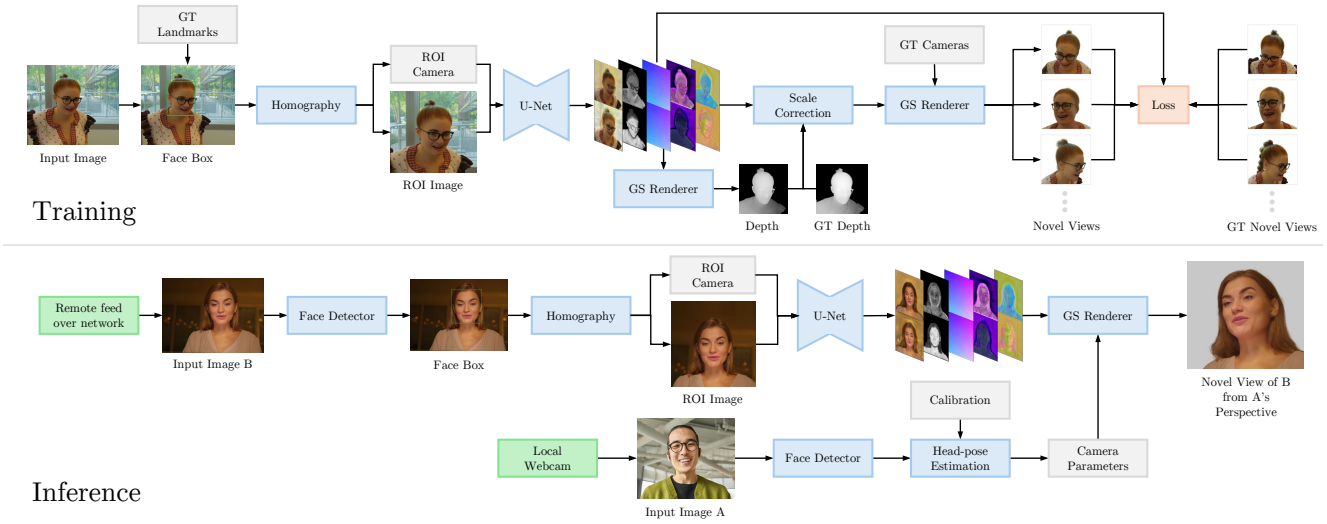


Figure 4. Overview of our the training (top) and inference (bottom) pipelines of our method.

where L_i is an image pair loss, U is the U-Net model and R is the differentiable Gaussian splat rendering function [10]. $N_v = 11$ is the number of views used in each sample for the supervision and N_s is the number of multi-view samples in the dataset. C_v are the camera parameters for view v and I_v^s is the ROI image for view v in sample s. $v = 1$ is taken as the input view. The image comparison loss, $L_i(X, Y) = L_e(X, Y) + \lambda_p L_p(X, Y)$, is composed of two terms: a Euclidean RGB distance term L_e and a perceptual loss term L_p with $\lambda_p = 0.5$. We composite the rendered image and the ground truth RGBA image over a random background colour for each sample.

This encourages the network to perform alpha matting by estimating the Gaussian transparency and subtracting the background colour from the predicted Gaussian colours.

Euclidean RGB distance. We compute L_e as the mean Euclidean distance in the the RGB space, i.e., $L_e(X, Y) = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} ||X_{i,j} - Y_{i,j}||_2$. This loss differs from the classical mean of absolute differences by being invariant to rotations in the 3D RGB colour space which helps reduce chromaticity aberration in the reconstruction, and is more tolerant to outliers than L_2 loss which helps to not over-penalize small silhouette misalignments.

Perceptual loss. We compute the perceptual loss L_p using the VGG network [29] at two resolutions: full (512×512) and half (256×256). We empirically found that the perceptual loss at half resolution helps to obtain better results qualitatively and quantitatively.

Regularization Losses. We extend the original Splatter Image [32] method by predicting multiple Gaussians per pixel. With this modification we often encounter local minima during training where only one 'layer' of Gaussians is used, even though this does not provide the best results. We therefore introduce losses on the splatter image opacities to avoid falling into poor local minima by penalizing cases where the average opacity is near 0 in any of the layers. Denoting $\sigma_{i,j,k}$ the opacity of the k^{th} Gaussian represented at coordinate (i, j), we compute mean opacities for Gaussians layer k as $\bar{\sigma}_k = \frac{1}{HW} \sum_{i,j} \sigma_{i,j,k}$. We penalize the means $\bar{\sigma}_k$ near zero using a loss $L_m = \frac{1}{K} \sum_k e^{-\tau \bar{\sigma}_k}$ with weight $\lambda_m = 5$ and $\tau = 50$, where K is the number of Gaussians per pixel. We also encourage the prediction of opaque Gaussians by using an opacity loss $L_\sigma = \frac{1}{K} \sum_k (1 - \bar{\sigma}_k)$ with weight $\lambda_\sigma = 10^{-4}$. This term acts as a small bias and helps in making the best use of the allowed number of Gaussians. Finally, we add a regularization term $L_c = \log(s)^2$ on the scale correction, s, obtained in the third step of the training loop to encourage the network to predict valid metric depths, with weight $\lambda_c = 1$. This helps significantly with convergence during training.

Jitter Loss. Temporal stability is important in the context of videoconferencing; flicker can be very distracting and degrade the experience. In order to improve the temporal stability of our prediction we use an approach similar to Zheng et al. [42]. Each sample, s, in a training batch is duplicated and random perturbations added to the face box, independently for each sample, to form sample s'. We then encourage stability by penalizing the difference in the renders between each pair of duplicated samples, $L_j = \frac{1}{N_v N_s} \sum_{v=1}^{N_v} \sum_{s=1}^{N_s} \|R(U(I_1^s), C_v) - R(U(I_1^{s'}), C_v)\|^2$. This encourages the DNN to favour consistency in the prediction over accuracy, and we can control the trade-off with weight $\lambda_j = 1.0$.



Figure 5. Snapshot from our prototype 3D video conferencing application. See supplementary video for details.

3. 3D Video Calls on Commodity Hardware

This section describes the live 3D video conferencing prototype application we use to demonstrate our method. Our application exploits the 'magic window' effect [14] to create the illusion of 3D on a 2D display. We have two subjects, A and B, both equipped with a webcam, a 2D computer screen and a GPU. Given the symmetry of the setup, we describe only what happens on subject A's machine, the bottom of Fig. 4 provides a graphical representation. The camera-screen relative pose is assumed known through calibration.

We render the appropriate new view of B's face on A's screen through the steps: (1) Detect A's face using an SSD-based detector [17] to get a bounding box, smooth it temporally to reduce jitter, and estimate the position of A's head assuming a fixed distance to the camera. (2) Detect B's face in the video feed received from B, and extract the ROI around B's face using a homography, as described in Sec. 2.1. (3) Predict a Gaussian splatter image in the ROI camera using the model described in Sec. 2.3. (4) Transform the Gaussians back to the original input camera coordinate system for B's video and render the 3D reconstruction of B in the view of subject A based on their estimated head position.

We run the prototype on a laptop with a NVIDIA 4090 Mobile GPU and video sent over a standard network connection. The application runs at 28FPS, a snapshot of the prototype in action is shown in Fig. 5, further demonstration is provided in the supplementary video.

4. Experiments and Results

To validate our approach, we compare our method to recent works across a number of benchmarks, for ablation experiments please consult the supplementary material. Note that we limit comparison to reconstruction methods which satisfy our requirement of *authenticity*, i.e., those which do not rely on an enrolled representation.

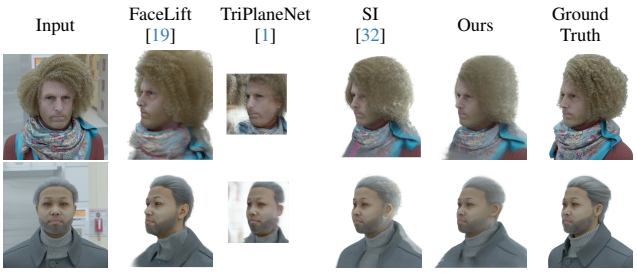


Figure 6. Qualitative comparison for images from the Cafca dataset [2]. Splatter Image (SI) has been retrained on our data.

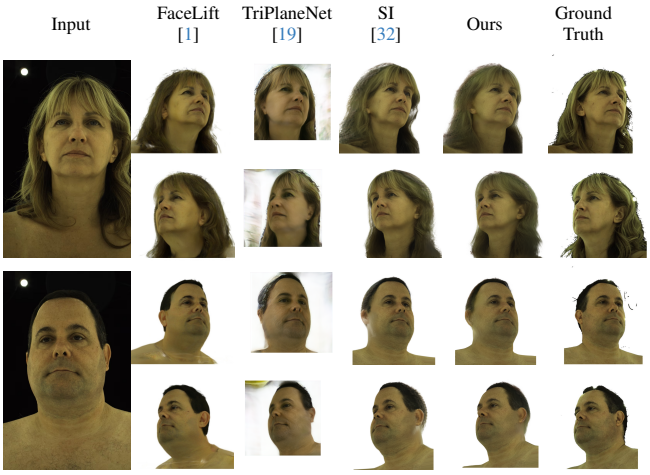


Figure 7. Qualitative comparison for images from the Ava-256 dataset [20]. Splatter Image (SI) has been retrained on our data.

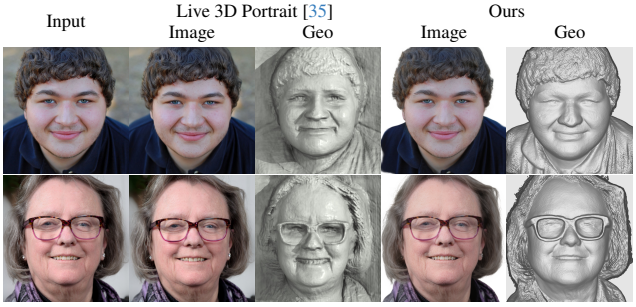


Figure 8. Qualitative comparison for images from the FFHQ dataset [9] showing reconstructed input images, and visualized geometry.

4.1. Experimental Setup

We use three datasets for our evaluation. Firstly, the Cafca dataset [2] which contains multi-view renders of synthetic humans with variation in identity, expression, clothing and environment. This dataset provides a diverse range of settings which emulate an in-the-wild context, but with limitations on visual quality and realism. Secondly, the Ava-256 dataset [20] contains multi-camera captures of people in a studio with controlled clothing and lighting. This dataset



Figure 9. Qualitative comparison for failure case of Live 3D Portrait (L3DP) [35]; our method is not constrained by EG3D [4], so can authentically represent all details from the input image, and can also create disconnected 3D structures like glasses frames.

provides less diversity than Cafca, but is not synthetic and so provides far higher realism. Finally, the FFHQ dataset [9] includes frontal, in-the-wild photos of a variety of people.

For Ava-256 and Cafca we follow the protocol of Lyu et al. [19] to evaluate novel view rendering (see supplementary material for details). We use PSNR, SSIM, LPIPS and DreamSim metrics to evaluate accuracy of image reconstruction, and ArcFace to assess identity preservation in novel views. For FFHQ we follow the protocol of Trevithick et al. [35] to evaluate input view reconstruction for a random subset of 500 images. We use LPIPS, DISTS and SSIM, but omit pose and identity metrics due to insufficient details to enable re-implementation.

4.2. Results

Results for novel view rendering on the Cafca dataset [2] are provided in Tab. 1. Our method outperforms others both quantitatively and qualitatively, and the results are of high quality. Compared to FaceLift [19] our results more accurately preserve the identity of the individual, although the visual quality of our results can degrade more significantly for extreme side views, as shown in Fig. 6. Overall, this produces numerically superior results, and is acceptable for our application where the viewing angle is limited. Results for novel view rendering on the Ava-256 dataset [20] are provided in Tab. 2. Results show a similar trend to those for Cafca, with superior identity preservation resulting in better overall metrics. See Fig. 7 for qualitative results. Qualitative results for in-the-wild images are shown in Figs. 1 and 10 and the supplementary material. Our method performs well across a diverse range of individuals and environments.

Results for input view reconstruction on the FFHQ dataset [9] are provided in Tab. 3. Our method does not perform as well as Live 3D Portrait [35] for DISTS, but is state-of-the-art for LPIPS and SSIM. Qualitative results shown in Fig. 8 demonstrate that the perceived difference in quality is minimal, and that our geometry is smoother and more accurate. Fig. 9 also shows some failure cases of Live 3D Portrait [35] which our method is able to address, such as fine details which are not in the EG3D [4] latent space and disconnected structures such as glasses frames which their approach is not able to represent. Note that our method

Table 1. Numerical results on Cafca [2] subset, following FaceLift protocol [19]. Splatter Image [32] is retrained on our data.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	DreamSim ↓	ArcFace ↓
Era3D [15]	13.69	0.7230	0.3662	0.2892	0.2978
TriPlaneNet [1]	15.82	0.7705	0.5303	0.1339	0.2826
LGM [33]	16.52	0.7933	0.3060	0.1552	0.2557
FaceLift [19]	16.61	0.7968	0.2694	0.1096	0.1573
Splatter Image [32]	20.04	0.8390	0.2244	0.0276	0.1202
Ours	**20.39**	**0.8462**	**0.1868**	**0.0286**	**0.1129**

Table 2. Numerical results on Ava-256 [20] subset, following FaceLift protocol [19]. Splatter Image [32] is retrained on our data.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	DreamSim ↓	ArcFace ↓
Era3D [15]	14.77	0.7963	0.2538	0.2515	0.3721
TriPlaneNet [1]	14.32	0.6846	0.3659	0.1692	0.2795
LGM [33]	14.05	0.8136	0.2476	0.1496	0.3142
FaceLift [19]	16.52	0.8271	0.2277	0.1065	0.1871
Splatter Image [32]	18.49	0.8258	0.2593	0.0725	0.1458
Ours	**19.11**	**0.8298**	**0.1908**	**0.0715**	**0.1216**

Table 3. Quantitative evaluation using LPIPS, DISTS, SSIM on 500 FFHQ images [9]. Evaluated only using the foreground on 256×256 images. * Use a different subset of 500 images; we find the deviation between different subsets to be negligable.

	LPIPS↓	DISTS↓	SSIM↑
ROME* [11]	.1158	.1058	.8257
Live 3D Portrait* [35]	.0468	**.0407**	.8981
TriPlaneNet [1]	.2109	.1580	.7075
Splatter Image [32]	.0876	.0916	.9165
Ours	**.0413**	.0649	**.9456**

Table 4. Models trained for 100 epochs with and without stability loss L_j with metrics computed with 250 subjects in the Ava256 dataset (more detail in the supplementary material), and our method compared with the Splatter Image method trained for 100 epochs.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	Jitter ↓
Splatter Image	18.224	0.8651	0.2300	0.006353
Ours no L_j	18.949	08658	0.1581	0.005020
Ours	**19.069**	**0.8682**	**0.1497**	**0.004967**

is also able to seamlessly fill the 'holes' behind occluding objects such as glasses frames.

Stability is an important attribute for our method. We devise a jitter metric to compare our method to Splatter Image [32]. We use images from 250 subjects in the Ava256 dataset with 10 views and 2 frames per subject. We align the reconstruction and ground truth in the first frame as described by Lyu et al. [19] and keep the same 2D alignment transform for the second frame to better measure jitter. The jitter metric for view v at frame t is defined as $||(I^v_t - I^v_{t-1}) - (R^v_t - R^v_{t-1})||$. We average this over the views and subjects. The quantitative results in Tab. 4 demonstrate that the jitter loss, L_j, contributes to improved stability.

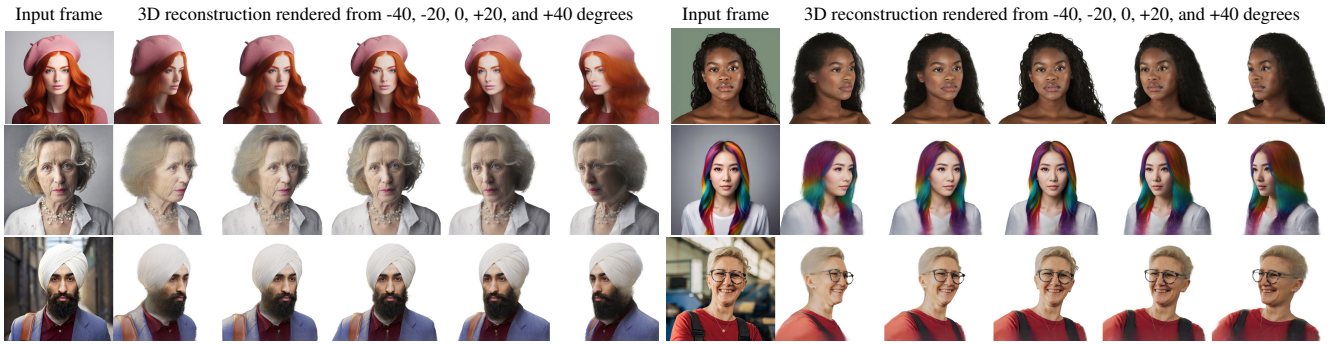


Figure 10. Qualitative results of our method. Our approach is able to accurately reconstruct headwear, glasses and other accessories, as well as capturing fine details of hair and skin.

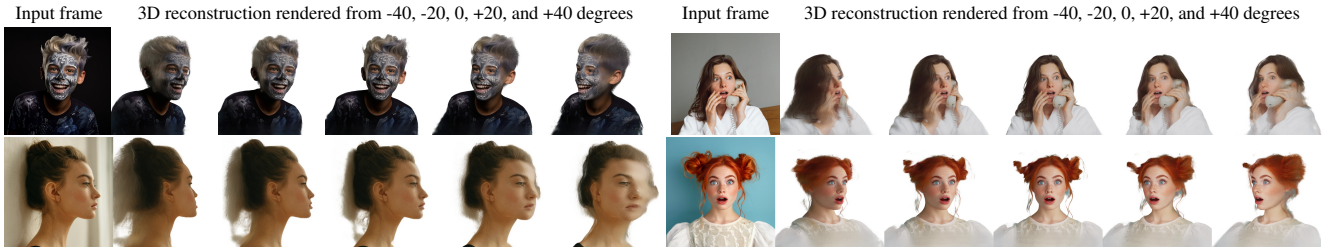


Figure 11. Limitations of our method include: unusual texture on the face, extreme input angles, hands/props, segmentation failures. We suspect these are due to lack of representation in the training data; future work will aim to improve the training data to address these issues.

4.3. Limitations

Our method has a number of limitations which largely derive from the training data. Hands, held objects and face paint are uncommon in the training data and the range of angles of the face and novel view cameras, relative to the input camera, are limited. As shown in Fig. 11, face-paint can result in hallucination of glasses frame structures. Extreme input angles can cause blurry/visually degraded predictions, although this can be desirable in our scenario to make clear that the prediction in this area is not based on directly observed data. Reconstruction quality of hands and held objects can be poor when viewed in novel views, and segmentation failures can also occur. Future work will aim to address these failures through improvements to the training data.

5. Potential societal impact

This work aims to bring the experience of remote communication closer to in-person communication, reducing geographic inequality. By relying only on commodity hardware, we avoid exacerbating inequality in technology access. Our method cannot be used to impersonate someone else in an online meeting, due to our requirement for authenticity, supporting trust in online communication and media. This work also supports authentic representation regardless of factors such as hair colour and type, skin tone, and choice of headwear, aiding diverse self-expression (see Fig. 10).

Like all human-centric computer vision, our 3D reconstruction may have lower accuracy for some demographic groups. We find that our use of synthetic data is helpful in addressing any lack of fairness we discover, given our precise control over the training data distribution. Nevertheless, there are aspects of human diversity that are not yet represented by our synthetic datasets (see Fig. 11), and there may also be lack of fairness that we have not yet discovered.

6. Conclusion

Online communication is growing ever more important, and yet 2D video meetings have challenges that are not solvable within the 2D paradigm. Recently there has been considerable interest in developing photorealistic avatars that can be controlled and animated from a broad range of signals. However, we argue that 3D video, with an emphasis on *authentic*, *realistic*, *live*, and *stable* reconstructions, is needed to support online meetings that bring us closer to in-person interactions, while preserving the properties that make videoconferencing so widely adopted today.

In this paper, we presented a method that fulfils the core properties required for authentic 3D video. There are many opportunities for improvement, and we hope to inspire future work that builds on these results. With this paper, we hope to highlight an open challenge for the community to demonstrate live 3D reconstruction from monocular inputs that can serve as a true analogue to 2D video, supporting the same diversity of expression that 2D video does today.

References

[1] Ananta R. Bhattarai, Matthias Nießner, and Artem Sevastopolsky. TriPlaneNet: An encoder for EG3D inversion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3055–3065, 2024. 2, 3, 6, 7

[2] Marcel C. Buehler, Gengyan Li, Erroll Wood, Leonhard Helminger, Xu Chen, Tanmay Shah, Daoye Wang, Stephan Garbin, Sergio Orts-Escolano, Otmar Hilliges, Dmitry Lagun, Jérémy Riviere, Paulo Gotardo, Thabo Beeler, Abhimitra Meka, and Kripasindhu Sarkar. Cafca: High-quality novel view synthesis of expressive faces from casual few-shot captures. In *SIGGRAPH Asia 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 2, 3, 6, 7

[3] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason Saragih. Authentic volumetric avatars from a phone scan. *ACM Trans. Graph.*, 41(4), 2022. 2

[4] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022. 2, 7

[5] Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao, and Yebin Liu. MonoGaussianAvatar: Monocular Gaussian point-based head avatar. In *ACM SIGGRAPH 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 2

[6] Stephan J. Garbin, Marek Kowalski, Virginia Estellers, Stanislaw Szymanowicz, Shideh Rezaeifar, Jingjing Shen, Matthew A. Johnson, and Julien Valentin. VolTeMorph: Realtime, controllable and generalizable animation of volumetric representations. *Computer Graphics Forum*, 43(6):e15117, 2024. 2

[7] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Learning neural parametric head models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[8] Charlie Hewitt, Fatemeh Saleh, Sadegh Aliakbarian, Lohit Petikam, Shideh Rezaeifar, Louis Florentin, Zafiirah Hosenie, Thomas J Cashman, Julien Valentin, Darren Cosker, and Tadas Baltrušaitis. Look Ma, no markers: holistic performance capture without the hassle. *ACM Transactions on Graphics (TOG)*, 43(6):#235:1–12, 2024. 4

[9] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 6, 7

[10] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 2, 5

[11] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *European Conference of Computer vision (ECCV)*, 2022. 7

[12] Tobias Kirschstein, Javier Romero, Artem Sevastopolsky, Matthias Nießner, and Shunsuke Saito. Avat3r: Large animatable Gaussian reconstruction model for high-fidelity 3D head avatars, 2025. 2

[13] Jason Lawrence, Dan B Goldman, Supreeth Achar, Gregory Major Blascovich, Joseph G. Desloge, Tommy Fortes, Eric M. Gomez, Sascha Häberling, Hugues Hoppe, Andy Huibers, Claude Knaus, Brian Kuschak, Ricardo Martin-Brualla, Harris Nover, Andrew Ian Russell, Steven M. Seitz, and Kevin Tong. Project Starline: A high-fidelity telepresence system. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 40(6), 2021. 1, 2

[14] Johnny Chung Lee. Head tracking for desktop virtual reality displays using the Wii remote, 2007. http://youtu.be/Jd3-eiid-Uw. 6

[15] Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wei Xue, Wenhan Luo, Ping Tan, Wenping Wang, Qifeng Liu, and Yike Guo. Era3D: High-resolution multiview diffusion using efficient row-wise attention. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 7

[16] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *Computer Vision – ECCV 2022*, pages 590–606, Cham, 2022. Springer Nature Switzerland. 4

[17] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016. 6

[18] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Trans. Graph.*, 40(4), 2021. 2

[19] Weijie Lyu, Yi Zhou, Ming-Hsuan Yang, and Zhixin Shu. FaceLift: Single image to 3D head with view generation and gs-lrm. *arXiv preprint arXiv:2412.17812*, 2024. 2, 6, 7

[20] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venshtain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matt Andromalos, Joanna Hsu, Thomas Dauer, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia

Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. *NeurIPS Track on Datasets and Benchmarks*, 2024. 6, 7

[21] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *The European Conference on Computer Vision (ECCV)*, 2020. 2

[22] Gary M. Olson and Judith S. Olson. Distance matters. *Hum.-Comput. Interact.*, 15(2):139–178, 2000. 1

[23] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. GaussianAvatars: Photorealistic head avatars with rigged 3D Gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 2

[24] Alexander Raake, Markus Fiedler, Katrin Schoenenberg, Katrien De Moor, and Nicola Döring. Technological factors influencing videoconferencing and Zoom fatigue. *arXiv:2202.01740 [cs]*, 2022. 1

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 3

[26] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable Gaussian codec avatars. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 130–141, 2024. 1, 2

[27] Jack Saunders, Charlie Hewitt, Yanan Jian, Marek Kowalski, Tadas Baltrusaitis, Yiye Chen, Darren Cosker, Virginia Estellers, Nicholas Gydé, Vinay P Namboodiri, et al. Gasp: Gaussian avatars with synthetic priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 271–280, 2025. 2, 3

[28] Martin Schuessler, Luca Hormann, Raimund Dachselt, Andrew Blake, and Carsten Rother. Gazing heads: Investigating gaze perception in video-mediated communication. *ACM Trans. Comput.-Hum. Interact.*, 31(3), 2024. 1

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 5

[30] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2, 3

[31] Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, Joao F Henriques, Christian Rupprecht, and Andrea Vedaldi. Flash3D: Feed-forward generalisable 3D scene reconstruction from a single image. *arXiv preprint arXiv:2406.04343*, 2024. 3

[32] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3D reconstruction. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 4, 5, 6, 7

[33] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. LGM: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 7

[34] John C. Tang, Kori Inkpen, Sasa Junuzovic, Keri Mallari, Andrew D. Wilson, Sean Rintel, Shiraz Cupala, Tony Carbary, Abigail Sellen, and William A.S. Buxton. Perspectives: Creating inclusive and equitable hybrid meeting experiences. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2), 2023. 1

[35] Alex Trevithick, Matthew Chan, Michael Stengel, Eric R. Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. Real-time radiance fields for single-image portrait view synthesis. In *ACM Transactions on Graphics (SIGGRAPH)*, 2023. 2, 3, 6, 7

[36] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and Baining Guo. Rodin: A generative model for sculpting 3D digital avatars using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4563–4573, 2023. 3

[37] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. FlashAvatar: High-fidelity head avatar with efficient Gaussian embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[38] Dejia Xu, Ye Yuan, Morteza Mardani, Sifei Liu, Jiaming Song, Zhangyang Wang, and Arash Vahdat. AGG: Amortized generative 3D Gaussians for single image to 3D. *Transactions on Machine Learning Research*, 2024. 2

[39] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic Gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[40] Yuelang Xu, Lizhen Wang, Zerong Zheng, Zhaoqi Su, and Yebin Liu. 3D Gaussian parametric head model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2

[41] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. GS-LRM: Large reconstruction model for 3D Gaussian splatting. *European Conference on Computer Vision*, 2024. 2

[42] Stephan Zheng, Yang Song, Thomas Leung, and Ian J. Goodfellow. Improving the robustness of deep neural networks via stability training. *CoRR*, abs/1604.04326, 2016. 5

[43] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4574–4584, 2022. 2

[44] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10324–10335, 2024. 2