

# Rig3DGS: Creating Controllable Portraits from Casual Monocular Videos

## -Supplementary-

Alfredo Rivero\*  
Stony Brook University  
arivero@cs.stonybrook.edu

Zhixin Shu  
Adobe Research  
zshu@adobe.com

ShahRukh Athar\*  
Stony Brook University, Captions  
sathar@cs.stonybrook.edu

Dimitris Samaras  
Stony Brook University  
samaras@cs.stonybrook.edu

## 1. Supplementary Video

In the file `Supplementary_Video.mp4` we provide results of driving Rig3DGS and RigNeRF (prior work and *only* other method that models both the human and the scene they're in) using a 3DMM with both constant and changing views. For reanimation, we use the expression and pose parameters from the driving 3DMM or video but keep the shape parameters unchanged. As can be seen, the photorealism of Rig3DGS's renders is far better than that of RigNeRF. Further, Rig3DGS is able to animate the portrait with high fidelity to the facial expression and head-pose of the driving 3DMM and video while simultaneously being view consistent.

## 2. Qualitative Comparison on Novel view synthesis

In this section, we provide a qualitative comparison between Rig3DGS and RigNeRF on novel views and facial expressions as mentioned in Section 4.3.2 of the paper. As can be seen, Rig3DGS generates significantly sharper renders for all subjects across various views and facial expressions.

## 3. Ablation of $\eta$ and $T(\cdot)$ of the deformation

In this section we ablate the importance  $\eta$  and  $T$  from Eq. 4 of the paper. We measure the PSNR, SSIM and LPIPS of full scene renders of both settings 1 and 2. As discussed in section 3.2 of the paper,  $\eta$  aids Rig3DGS in modeling small deformations that lie outside the subspace defined by the deformation of vertices that are bound to each gaussian while  $T$  aids in modeling small movements of the body. This is quantitatively verified in Table 1, where we see that without  $\eta$  and  $T$  the reconstruction quality drops significantly. Using  $T$  helps in modeling body motions and thus leads to

better reconstruction while using both  $\eta$  and  $T$  (labeled as 'Full Deformation') yields the best results.

Full Scene Metrics	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
w/o $\eta$ and $T$	25.74	0.8273	0.1339
w/o $\eta$	25.98	0.8583	0.1288
Full Deformation	26.16	0.8797	0.1292

Table 1. Ablation of  $\eta$  and  $T$  deformation model.

## 4. Ablation of number of nearest neighbors bound to each gaussian

In this section we ablate the number of nearest neighbors that we bind to each gaussian to create the deformation subspace. As in the previous section, we measure the PSNR, SSIM and LPIPS of full scene renders of both settings 1 and 2. From Table 2, we see that both  $K = 10$  and  $K = 20$  give similar PSNR and SSIM values and  $K = 5$  performs much worse. In contrast, the LPIPS value for  $K = 5, 10$  and  $20$  are about the same. Thus, in order to get the best quality and save computational resources we use  $K = 10$  for all experiments in the main paper and supplementary videos.

Full Scene Metrics	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
$K = 5$	25.89	0.8543	0.1295
$K = 10$	26.16	0.8797	0.1292
$K = 20$	26.13	0.8814	0.1312

Table 2. Ablation of nearest vertices used to define the deformation subspace.

## 5. Limitations

Like prior work in facial reanimation, Rig3DGS struggles in modeling high-frequency illumination dependent effects

\*Equal contribution



Figure 1. Qualitative comparison of Subjects 1-6 in Setting 2. As can be seen, Rig3DGS generates higher quality renders than RigNeRF across all subjects.



Figure 2. Like prior work, Rig3DGS struggles in modeling high-frequency illumination dependent effects such as cast-shadows.

such as cast-shadows. This is due to the inability of spherical harmonics to adequately learn high-frequency illumination effects. We believe that modeling such illumination is a fruitful area of future research.