GGAvatar: Dynamic Facial Geometric Adjustment for Gaussian Head Avatar

Supplementary Material

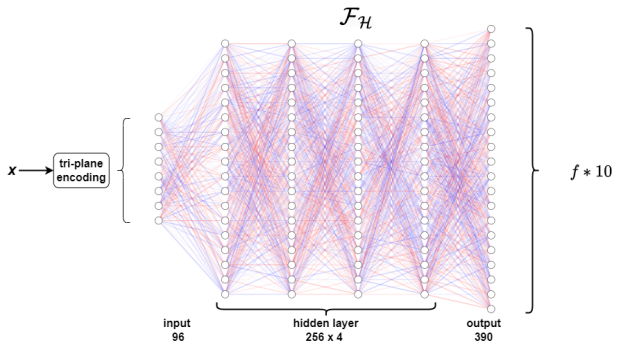

Figure 1: *High-frequency deformation decoder MLP architecture*

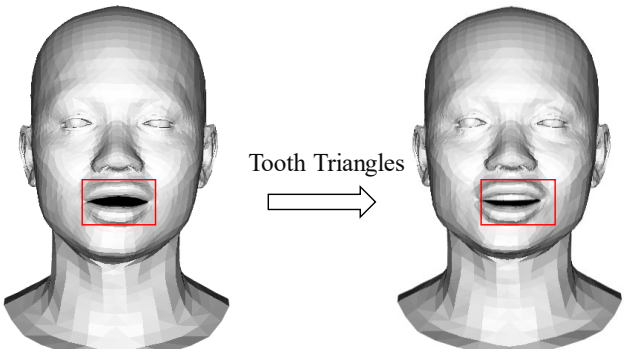

Figure 3: *Additional triangles for intraoral teeth representation*

1. Implementation Details

1.1. Network Architecture

In Figure 1 and Figure 2, we show our network architectures. We use an MLP with four Fully Connected layers (FC) containing 256 neurons as the high-frequency deformation decoding network. The input to this network is the tri-plane encoding of global Gaussian coordination, as shown in Figure 1. We use an MLP with four fully connected layers containing 128 neurons as the facial parameters encoding network. The input to this network comprises the expression parameters, jaw pose parameters, eyelids parameters, and eyes pose parameters from the FLAME model, as shown in Figure 2.

Table 1: *Additional ablation study on ID 1.*

Methods	PSNR↑	SSIM ↑	LPIPS ↓
w/o $\mathcal{L}_{\text{position}}$ and $\mathcal{L}_{\text{scaling}}$	31.92	0.943	0.067
w/o perceptual loss	32.05	0.944	0.086
Ours	32.00	0.944	0.068

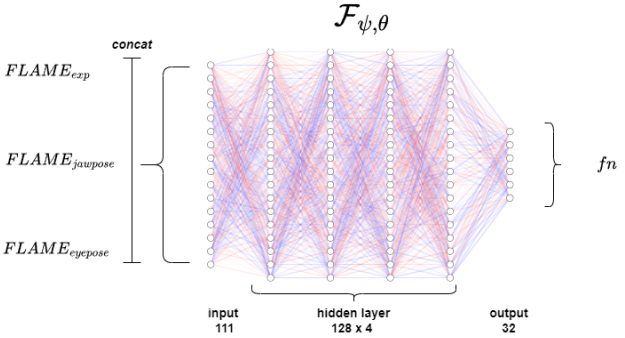

Figure 2: *Facial parameters encoder MLP architecture*

1.2. Additional Flame Meshes

Due to the limitations of the vanilla FLAME mesh in capturing intraoral details, we have added 168 mesh triangles to the FLAME template to construct the teeth, as illustrated in Figure 3. Furthermore, we link the rigid transformation of the upper and lower tooth triangles to the camera pose and the jaw joint, respectively. This improves our avatar's fidelity and prevents teeth and lips from sticking together due to the Gaussian function modeling intraoral details in the lip region, as shown in Figure 4.

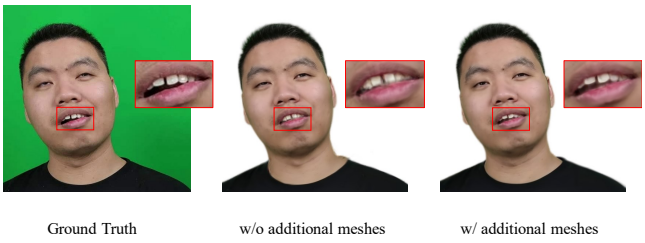

Figure 4: *Incorporating triangles that move rigidly with the head and jaw helps Gaussian Primitives accurately capture intraoral details.*

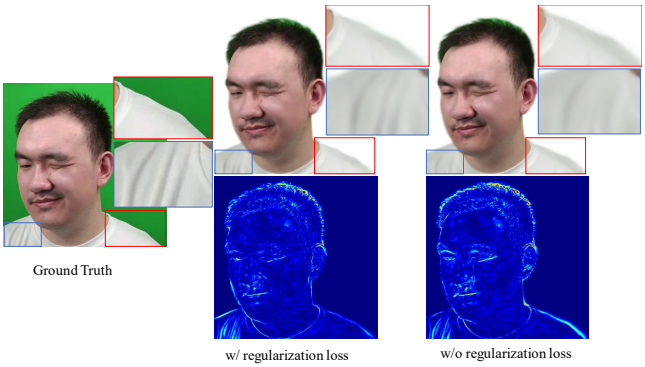

Figure 5: *The position and scaling losses help prevent artifacts and jitter during animation with novel expressions and poses.*

2. Additional Ablations and Results

2.1. Additional Ablations

Regularization Loss

To alleviate unnatural jitter in the synthesized video, we regularize the position and scaling of the Gaussians. To demonstrate the effectiveness of this approach, we conduct additional ablation studies, as shown in Table 1. Without constraints on position and scale, Gaussian primitives can freely move and scale in space to minimize rerendering errors, which means that removing these constraints does not significantly degrade metrics. However, unconstrained Gaussians tend to overfit the training frames, resulting in unnatural artifacts on a visual level and jitter when new expressions are introduced, as illustrated in Figure 5. Therefore, regularizing Gaussian properties is essential to enhance their robustness for animation.

Perceptual Loss

In addition to supervising pixel-level errors, we also employ perceptual loss to enhance the photorealism of the generated images. As shown in Table 1, ablating the perceptual loss results in a significant increase in perceptual error. Figure 6 presents a visual comparison of results with/without perceptual loss. As discussed, incorporating perceptual loss helps in recovering finer facial details.

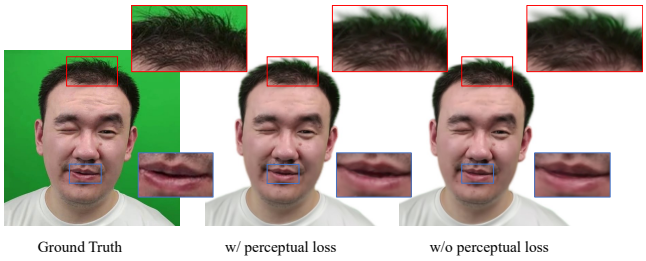

Figure 6: *Perceptual loss aids in reconstructing fine-grained facial attributes.*

2.2. Additional Results

In Figure 7, we show the error map, which can better demonstrate the accuracy of the reconstructions, we calculated the L2 error maps for each reconstructed image to compare the quality of the reconstructions. From the figure, it is evident that our model achieves superior accuracy, exhibiting the lowest number of errors across all images.

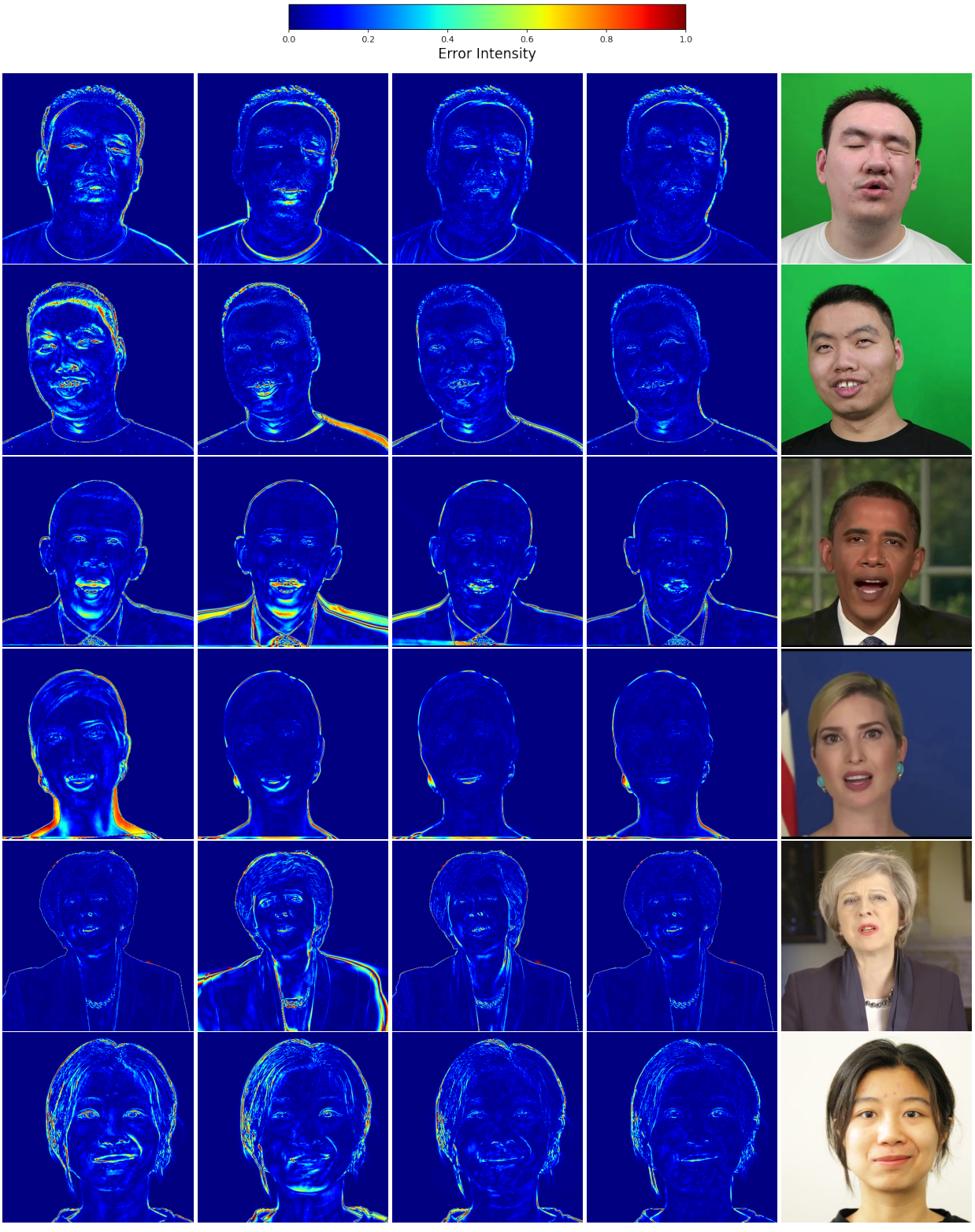

Figure 7: *Error map comparison, we use L2 distance.*