

LegacyAvatars: Volumetric Face Avatars For Traditional Graphics Pipelines

Supplementary Material

6. Implementation Details

Architecture details. The warp and texture predictors are implemented as 6-layer MLPs, where the subject-specific embeddings ϕ_w, ϕ_t are concatenated to the input coordinates and each of the learned latent codes in embedding matrices E_w, E_t are concatenated to the features after the third layer and fed through the model in parallel to produce a basis of warps and textures during each forward pass. In this formulation, variations across the elements of the bases are offloaded to single matrix, while the weights of the network are shared. This provides sufficient variability within the bases that can generate deformations and appearances with high expressivity, while also maintaining computational efficiency at training time.

3DMM and fitting. Our 3DMM includes linear bases of identity and expression. For each frame, we fit the 3DMM by estimating 599 probabilistic landmarks in 2D and optimizing identity, expression, rotation, and translation parameters of the 3DMM using a loss function that encourages consistency between per-vertex landmarks of the 3DMM and the observed 2D landmarks [55]. The parameter size of our expression model is $p = 63$.

Synthetic data. We use the synthetic face dataset introduced in [6], where we augment a subset of the subjects in this dataset to the real data. During training, we construct our batches by gathering 50% of the rays from the real subject and the other 50% from synthetic subjects. To ensure broad coverage of facial dynamics, synthetic expressions are drawn uniformly across the entire dataset. We illustrate the effectiveness of our joint real–synthetic training in Fig. 10, where we observe that in the absence of synthetic data, our model is prone to geometric instabilities and may fail to generalize to novel expressions.

Exported assets. The blend weights for our warp and texture bases can be efficiently computed at rendering time by linearly mapping $p = 63$ dimensional expression coefficients to $W = T = 12$ coefficients. Including the learned constant offset in this mapping, this results in two weight matrices of size 12×64 .

Our canonical UV values, warp basis, and texture basis are all exported in the UV space, where the resolution and the precision can be modified for different application needs. Please refer to Fig. 11 for visualizations of our assets. For renders at 0.5K resolution, we found that mesh, warp map, and texture map resolutions of 512×512 are sufficient to preserve the overall visual quality. Here, a 32-bit precision is maintained for UV values, while the appearance

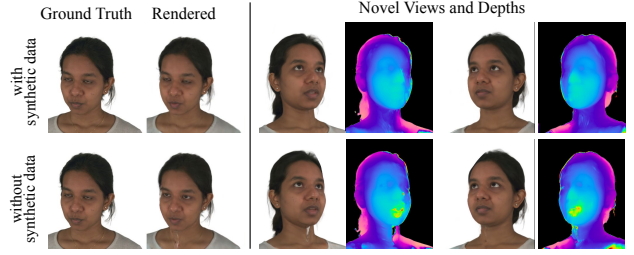


Figure 10. **Ablation on synthetic data.** The stability and over-fitting challenges can be mitigated by introducing synthetic face data jointly trained with the real subject. This helps with generalization to novel expressions in addition to regularization of the learned face geometry.

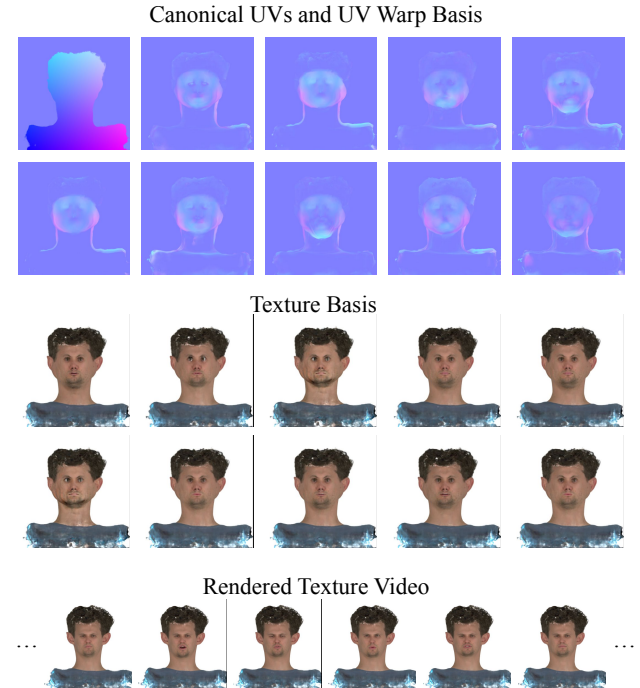


Figure 11. **Visualizations of the assets.** Illustrating a subset of the learned warps and appearances. With tracked expression coefficients of a 3DMM, these assets can be used to render a texture video shown at the bottom. All images are alpha-composited for visualization purposes.

is exported as 8-bit RGBA maps, where the view-dependent radiances are discarded.



Figure 12. **Comparisons on novel view synthesis.** Our model can synthesize novel views at a comparable visual quality to MonoAvatar++ [3] and GaussianAvatars [41], while being less prone to floater artifacts in NeRFs, and inherently preventing primitive ordering artifacts in 3DGS-based methods.

Table 2. **Ablation study on sizes of warp and texture bases.** Texture and warp basis sizes improve rendering quality and generalization. These metrics are obtained on cropped images that include the face region only.

	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
$W = T = 16$	29.67 ± 1.36	0.897 ± 0.012	0.258 ± 0.017
$W = T = 8$	29.47 ± 1.39	0.894 ± 0.012	0.259 ± 0.017
$W = T = 4$	29.38 ± 1.41	0.892 ± 0.012	0.263 ± 0.016
$W = T = 2$	28.46 ± 1.21	0.882 ± 0.012	0.276 ± 0.019

7. Additional Results and Comparisons

Novel view synthesis comparisons. We provide comparisons on novel view synthesis with the state of the art methods, see Fig. 12. NeRF-based methods like MonoAvatar++ [3] can manifest floating artifacts and 3DGS-based methods may result in popping-like artifacts due to explicit sorting of primitives [42]. Our method is not prone to such artifacts by design, and the exported textured meshes can be edited by an artist, providing additional flexibility to remove visual seams as a post-processing step. Please see the supplementary video for better visualizations.

Warp and texture basis size ablations. To provide more insights into our model, we evaluate its expressiveness by modifying its warp and texture basis sizes. We illustrate our results in Fig. 13 and evaluation metrics in Tab. 2, where we observe that the overall rendering quality suffers and the renders manifest artifacts as we reduce the basis sizes. Furthermore, the model does not generalize to novel facial expressions and eye gazes as we reduce its capacity.



Figure 13. **Ablation on sizes of warp and texture bases.** Sufficient number of blendable warps and textures is crucial to achieve good rendering quality and generalization to novel expressions.



Figure 14. **Limitations.** Layered mesh representations may suffer from *shell* artifacts at extreme angles. While our approach outperforms existing baselines on visual quality on novel views at moderate poses (*left*), at more extreme out-of-training profile views (*right*) it also suffers from *shell* artifacts similar to the baseline.

8. Limitations and Future Work

The basic premise of our representation is the discretization of scene components like geometry, appearance and deformations. While this enables traditional rendering, it also inherently limits the representational capacity compared to continuous volumes such as radiance fields and 3D Gaussians. Furthermore, since we project expression parameters onto low-dimensional blend weights, our model may exhibit blurring artifacts for extreme expressions, particularly those involving strong deformations around the mouth. We also note that our model does not explicitly account for the head pose, and hence the neck and torso regions may show slight instabilities in cases where the training sequences involve strong head pose variations.

Our layered mesh with transparency can be seen as a generalization of multiplane imaging (MPI) [63], where we instead learn a set of surfaces that follow a coarse face geometry and represent dynamic scenes. Due to such coarse

geometry, there exists a fundamental limit to the viewing angle range for artifact-free rendering [47]. At extreme angles, our representation can manifest shell artifacts, please see Fig. 14. We have also not tested our representation for controlling/animating large deformations such as head/neck rotations. These may also require additionally including and optimizing for a root joint UV-deformation of the layered mesh in the neck region. Optimizing and regularizing the topology of the layered mesh and the texture basis to improve the overall representational capacity could be an interesting future line of research.

The enrollment phase in our pipeline relies on ML training and inference, future work could explore simplifying this process. Recently, generative models have been used to learn a strong prior of face geometry and appearance [52], allowing direct regression of the face volume from even single images. Such quick and efficient techniques can further help reduce the compute and memory cost of the enrollment phase by learning a data-driven generative model. This can particularly help extend our representation to consumer use-cases such as monocular enrollment and tracking.

We believe that our work lays the important groundwork of building a novel representation that is capable of representing, animating and synthesizing volumetric effects in traditional graphics pipelines, and future work can build on and expand it towards even more practical settings.