Avat3r

Large Animatable Reconstruction Model for High-fidelity 3D Head Avatars

Tobias Kirschstein[1,2] - Javier Romero[2] - Artem Sevastopolsky[1,2] - Matthias Nießner[1] - Shunsuke Saito[2]

[1]Technical University of Munich
[2]Meta Reality Labs

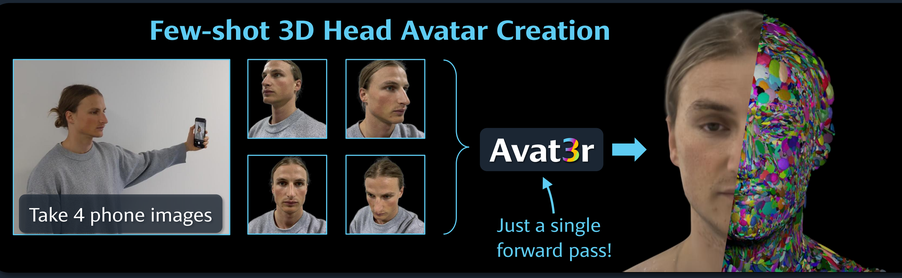
ICCV OCT 19-23, 2025 HONOLULU HAWAII

Project Page

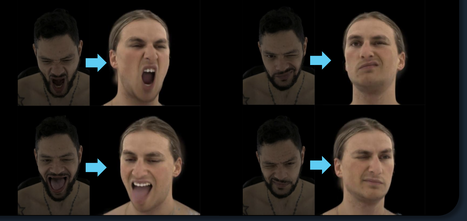
Highlights

- High-quality 3D Head Avatars from just 4 images
- Make Large Reconstruction Models (LRMs) animatable via simple cross-attention to the animation signal (facial expression codes)
- No reliance on mesh-based 3D face model: Learn facial animations entirely from data
- Improve per-pixel 3D Gaussian predictions with position maps from DUSt3R and feature maps from Sapiens

Few-shot 3D Head Avatar Creation

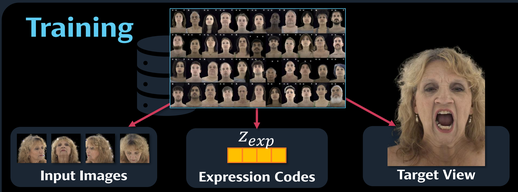


Expressive Animation



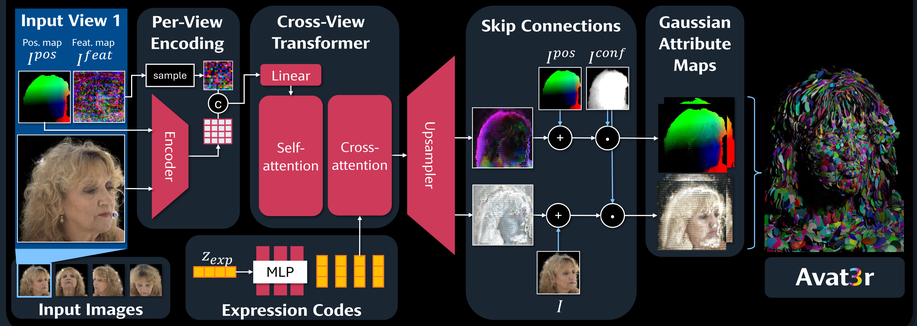
More few-shot Avat3rs

Training

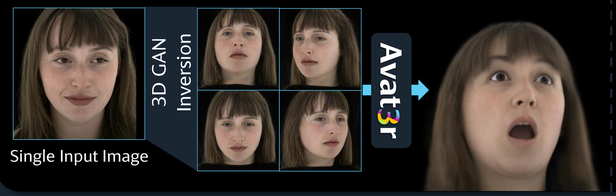


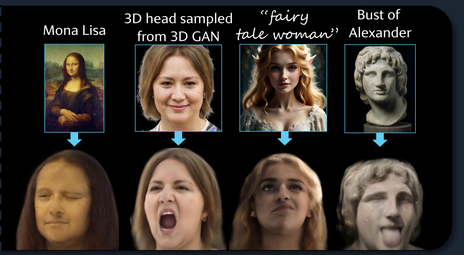
- Train on 244 persons from ava256 dataset (4TB)
- Predict 1 Gaussian for each pixel in the 4 input images

Method Overview



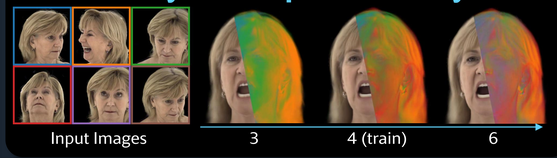
Single-image Pipeline





Comparisons

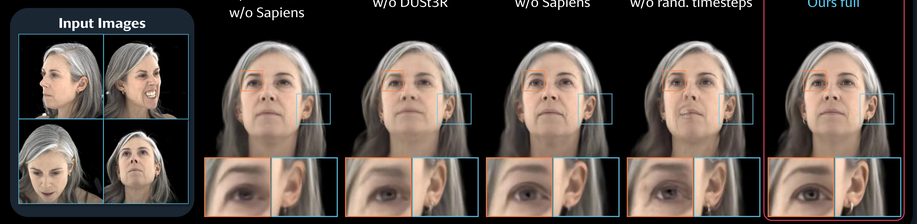
Analysis of Input Flexibility



Ablations

	$\mathcal{S}$	$\mathcal{D}$	$\mathcal{T}$	PSNR↑	LPIPS↓	AKD↓
Just cross-attention			☒	20.8	0.439	8.60
w/o Sapiens		☒	☒	20.9	0.434	8.59
w/o Dust3r	☒		☒	21.1	0.429	9.60
w/o rand. timesteps	☒	☒		21.3	**0.409**	8.86
Avat3r	☒	☒	☒	**21.6**	0.410	**8.08**

$\mathcal{S}$: Use Sapiens feature maps
$\mathcal{D}$: Use DUSt3R position maps
$\mathcal{T}$: Sample inconsistent expressions during training



Conclusion

- → Feed-forward 3D head avatar reconstruction from 4 images (no optimization needed)
- → Surprisingly good generalization from just 244 persons
- → Cross-attention is enough to model complex facial expressions
- → No 3D face model needed → Can animate tongue
- → Model is robust to inconsistent expressions in a phone scan