

Benchmarking and Analyzing 3D-aware Image Synthesis with a Modularized Codebase

Qiuyu Wang¹, Zifan Shi², Kecheng Zheng¹, Yinghao Xu³, Sida Peng⁴, Yujun Shen¹

¹Ant Group, ²HKUST, ³CUHK, ⁴ZJU



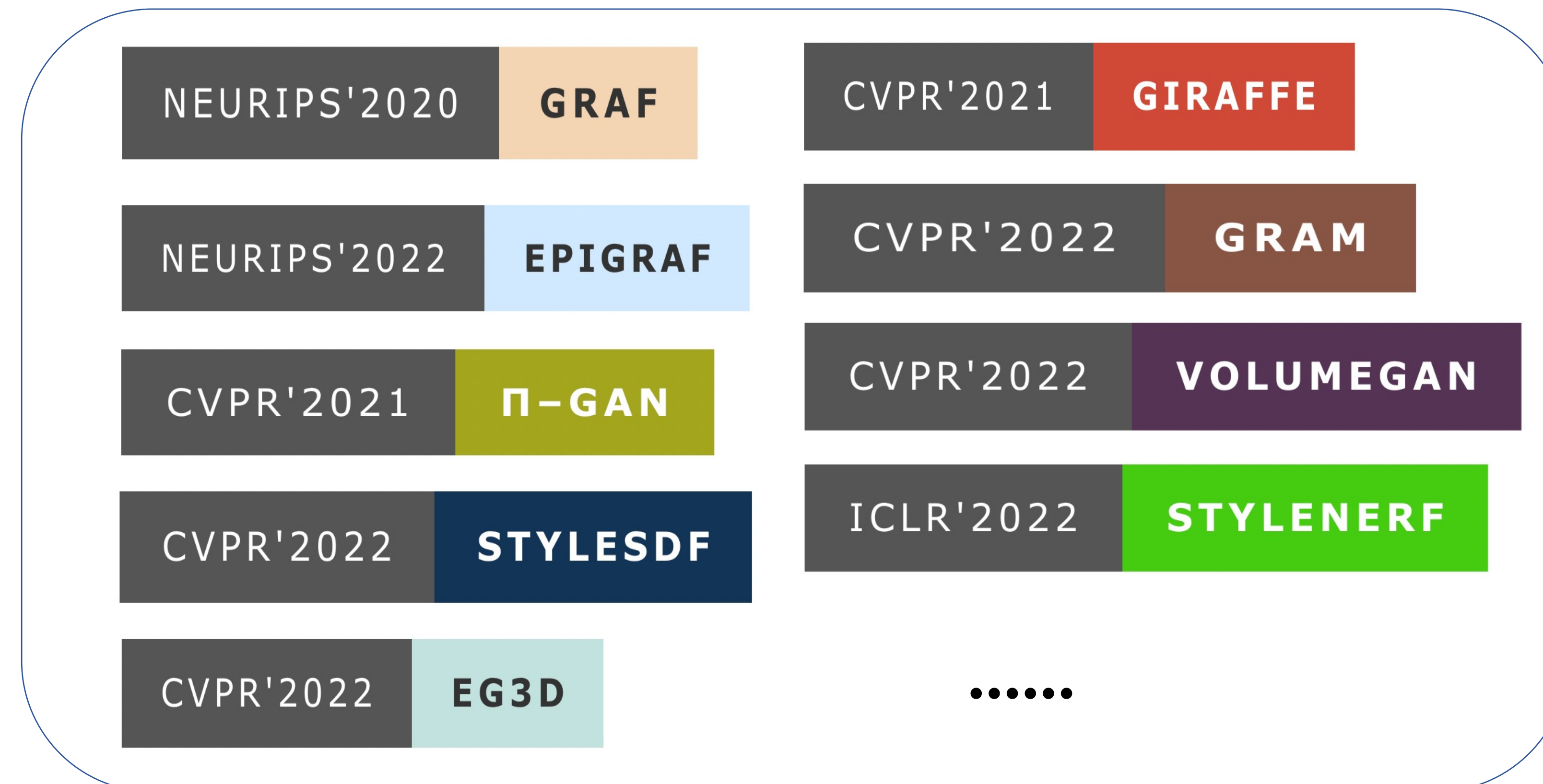
Outline

- Background & Motivation
- Our Modularized pipeline
- Experiment results & Analyses

Background & Motivation

Problem

3D-aware Image Synthesis Models



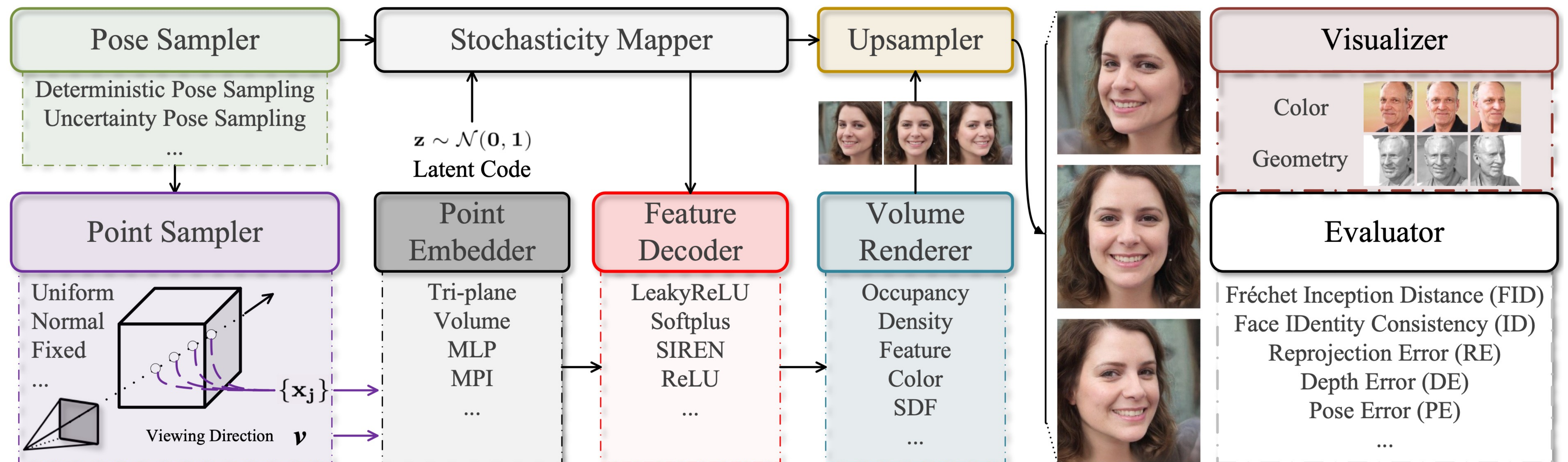
- 😞 Developed with different codebases
- 😞 Entangled implementation
- 😞 No unified and modularized framework

Our Modularized pipeline

Our Solution

- 😊 Build a highly-modularized easy-to-use codebase for 3D-aware image synthesis
- 😊 Allows users to replace a particular module arbitrarily and independently
- 😊 Perform a variety of in-depth analyses regarding different modules

Modularized pipeline for 3D-aware image synthesis

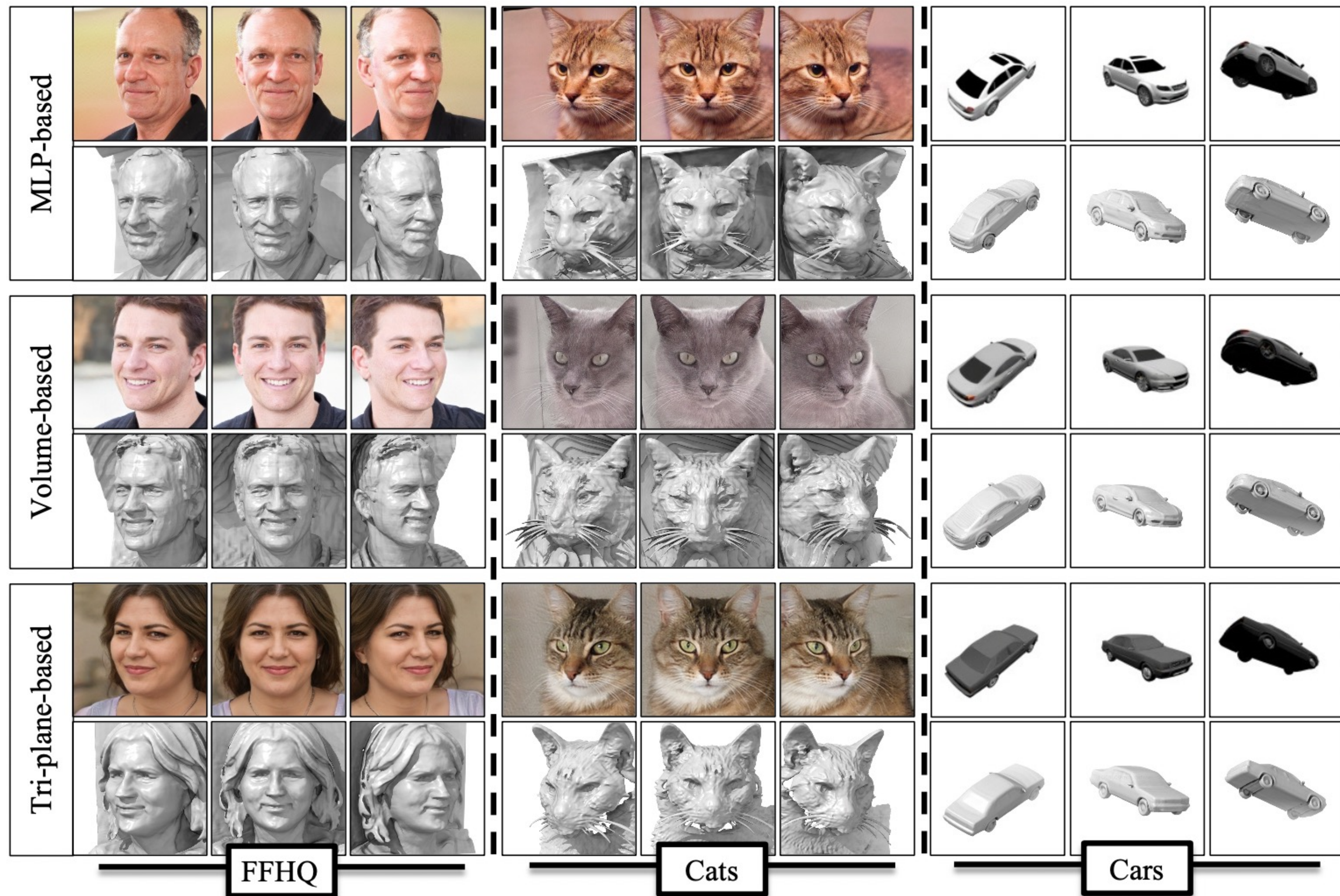


Experiment results & Analyses

Supported methods and reproduced results

Method	Pose Sampler	Point Embedder	Feature Decoder	Volume Renderer	Upsampler	Resolution	Official	Reproduction
GRAF [46]	Stochastic	MLP	ReLU	Density, Color	No	128×128	46.30	45.50
π -GAN [6]	Stochastic	MLP	SIREN	Density, Color	No	128×128	29.90	27.81
StyleSDF [40]	Stochastic	MLP	SIREN	SDF, Color, Feature	Yes	256×256	11.50	10.96
						512×512	10.07	10.71
						1024×1024	10.01	10.14
StyleNeRF [20]	Stochastic	MLP	ReLU	Density, Color, Feature	Yes	256×256	8.00	8.31
						512×512	7.80	7.37
						1024×1024	8.10	8.08
VolumeGAN [56]	Stochastic	Volume	LeakyReLU	Density, Color, Feature	Yes	256×256	9.10	10.37
GRAM [14]	Deterministic	MPI	SIREN	Occupancy, Color	No	256×256	14.50	13.83
EpiGRAF [50]	Deterministic	Tri-plane	LeakyReLU	Density, Color	No	512×512	9.92	9.19
EG3D [8]	Deterministic	Tri-plane	Softplus	Density, Color, Feature	Yes	256×256	4.80	4.72
						512×512	4.70	4.63

Point embedders



Qualitative comparison across various single point embedders on FFHQ, Cats and ShapeNet Cars

Point embedders

Point Embedder			FFHQ [26]					Cats [62]	Cars [10]
MLP	Volume	Tri-plane	FID↓	ID↑	DE↓	PE↓	RE↓	FID↓	FID↓
✓	✗	✗	5.15	0.777	0.470	$5.0e^{-4}$	0.091	4.05	2.42
✗	✓	✗	4.65	0.778	0.413	$5.1e^{-4}$	0.085	3.59	2.25
✗	✗	✓	4.72	0.743	0.547	$4.5e^{-4}$	0.111	3.99	2.75
✓	✓	✗	4.70	0.773	0.334	$5.1e^{-4}$	0.086	3.87	2.55
✓	✗	✓	4.69	0.748	0.465	$5.3e^{-4}$	0.104	4.42	2.59
✗	✓	✓	4.68	0.735	0.378	$4.6e^{-4}$	0.100	4.41	2.78
✓	✓	✓	4.62	0.769	0.467	$4.7e^{-4}$	0.091	4.70	2.65

- Different point features exhibit competitive capacities
- The contribution of multiple point features is marginal compared to a single type of point feature

Feature Decoder

Point Embedder	Depth	FFHQ [26]				
		FID↓	ID↑	DE↓	PE↓	RE↓
MLP	4	17.22	0.761	0.807	$12.2e^{-4}$	0.105
	8	7.39	0.782	0.552	$7.3e^{-4}$	0.087
	16	5.15	0.777	0.470	$5.0e^{-4}$	0.091
Volume	4	5.65	0.784	0.437	$4.4e^{-4}$	0.095
	8	5.18	0.787	0.381	$4.0e^{-4}$	0.100
	16	4.65	0.778	0.413	$5.1e^{-4}$	0.085
Tri-plane	2	4.72	0.743	0.547	$4.5e^{-4}$	0.111
	4	4.77	0.750	0.414	$4.4e^{-4}$	0.101
	8	5.58	0.750	0.566	$5.6e^{-4}$	0.108

Activation Type	FFHQ [26]				
	FID↓	ID↑	DE↓	PE↓	RE↓
- w/ upsampler	256 × 256				
SIREN	11.66	0.763	0.352	$9.1e^{-4}$	0.089
ReLU	7.39	0.782	0.552	$7.3e^{-4}$	0.087
- w/o upsampler	64 × 64				
SIREN	6.58	0.741	0.340	$6.6e^{-4}$	0.071
ReLU	7.30	0.729	0.498	$4.6e^{-4}$	0.084

- The depth only matters for MLP-based point embedder
- SIREN is better than ReLU when upsampler module is absent

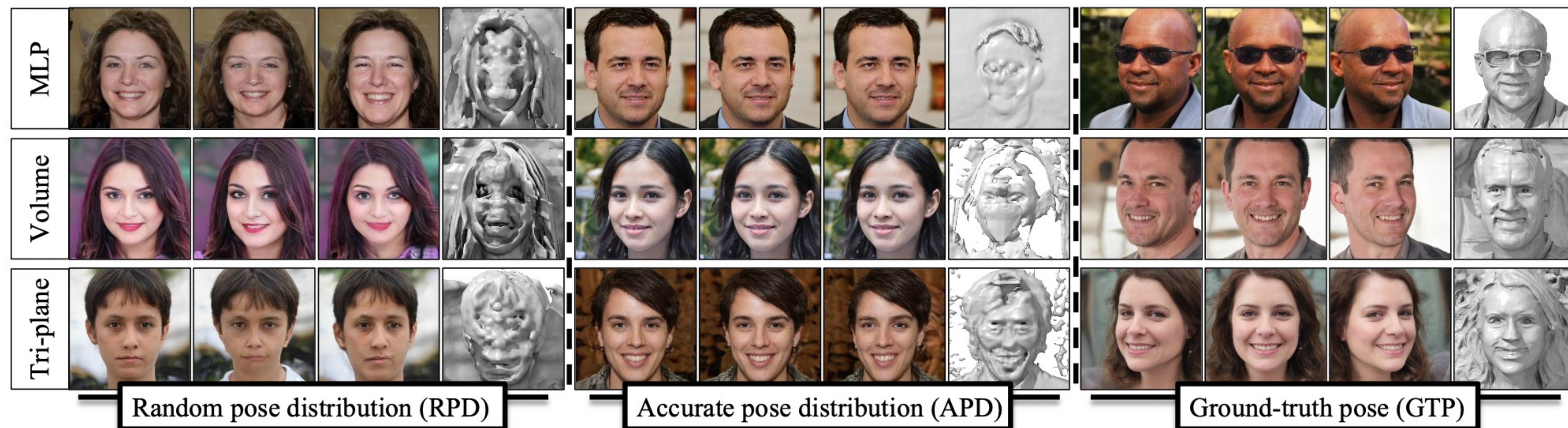
Geometric representation

Geometric Representation		FFHQ [26]				
		FID↓	ID↑	DE↓	PE↓	RE↓
MLP	SDF	8.87	0.610	0.874	$5.9e^{-4}$	0.184
	Density	5.15	0.777	0.470	$5.0e^{-4}$	0.091
Volume	SDF	7.27	0.676	0.938	$5.0e^{-4}$	0.200
	Density	4.65	0.778	0.413	$5.1e^{-4}$	0.085
Tri-plane	SDF	13.31	0.534	0.626	$10.9e^{-4}$	0.161
	Density	4.72	0.743	0.547	$4.5e^{-4}$	0.111

SDF-based representation currently lags behind the density-based one

Pose priors

Pose Prior	FFHQ [26]				
	FID↓	ID↑	DE↓	PE↓	RE↓
MLP w/ RPD	14.56	0.413	1.513	$5.8e^{-2}$	0.405
MLP w/ APD	9.96	0.788	1.659	$5.9e^{-2}$	0.407
MLP w/ GTP	5.15	0.777	0.470	$5.0e^{-4}$	0.091
Volume w/ RPD	10.47	0.429	1.562	$5.5e^{-2}$	0.390
Volume w/ APD	7.34	0.731	1.125	$4.8e^{-2}$	0.367
Volume w/ GTP	4.65	0.778	0.413	$5.1e^{-4}$	0.085
Tri-plane w/ RPD	15.18	0.427	2.181	$5.6e^{-2}$	0.379
Tri-plane w/ APD	5.45	0.764	1.502	$5.4e^{-2}$	0.405
Tri-plane w/ GTP	4.72	0.743	0.547	$4.5e^{-4}$	0.111



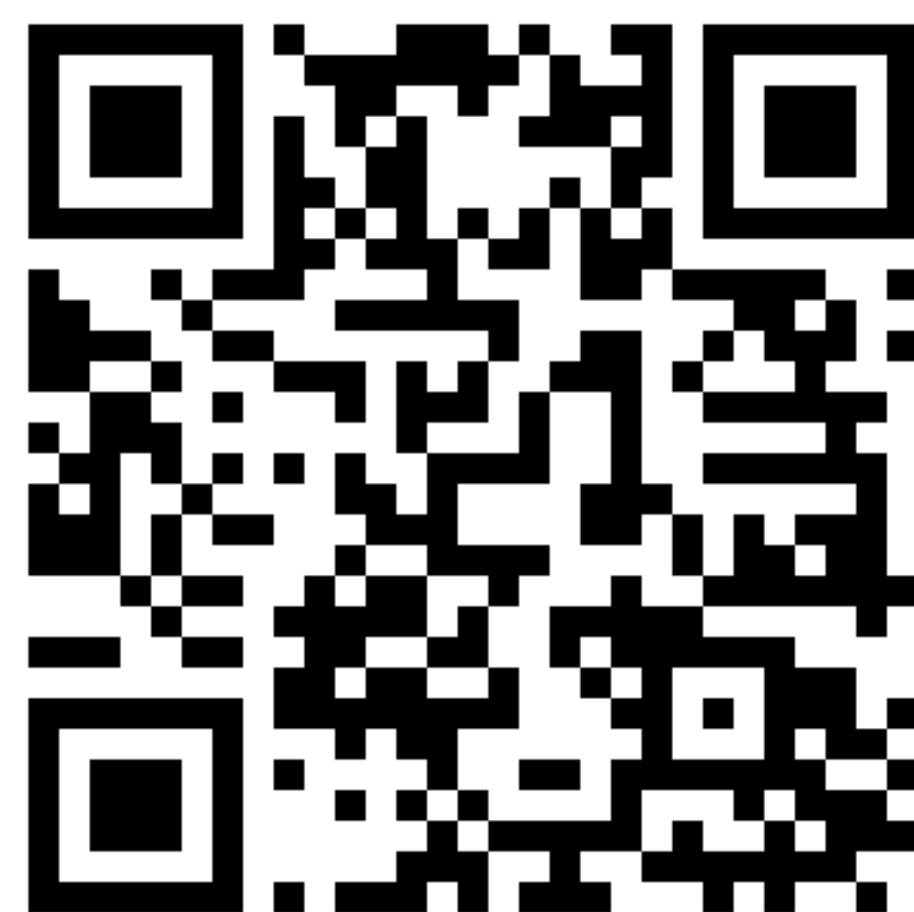
The more accurate the poses are, the better the generation quality is

Upsampler

Upsampler	Resolution	FFHQ [26]					Training Time	Inference Speed
		FID↓	ID↑	DE↓	PE↓	RE↓		
✓	256×256	4.72	0.743	0.547	$4.5e^{-4}$	0.111	2.7 Days	49 FPS
✗	256×256	6.86	0.749	0.443	$6.2e^{-4}$	0.104	6.9 Days	20 FPS

Upsamplers benefit the quality but harm the multi-view consistency

Thanks!



Paper



Code