3D-Aware Generative Adversarial Networks for High-resolution Image Synthesis

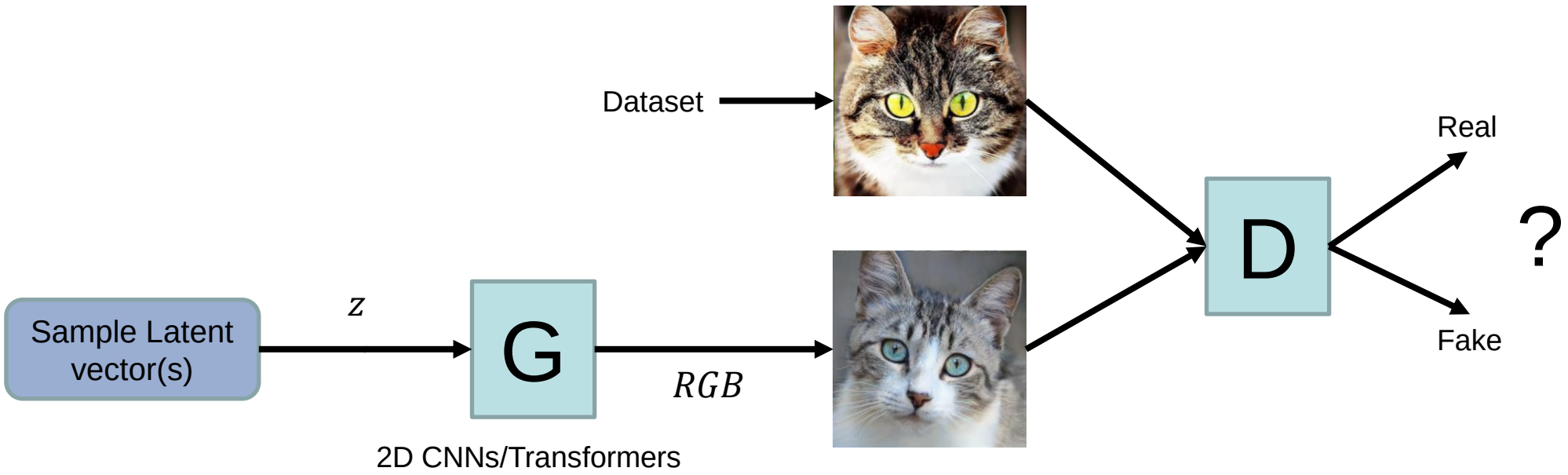
Jiatao Gu

Meta (FAIR Labs)

Motivation

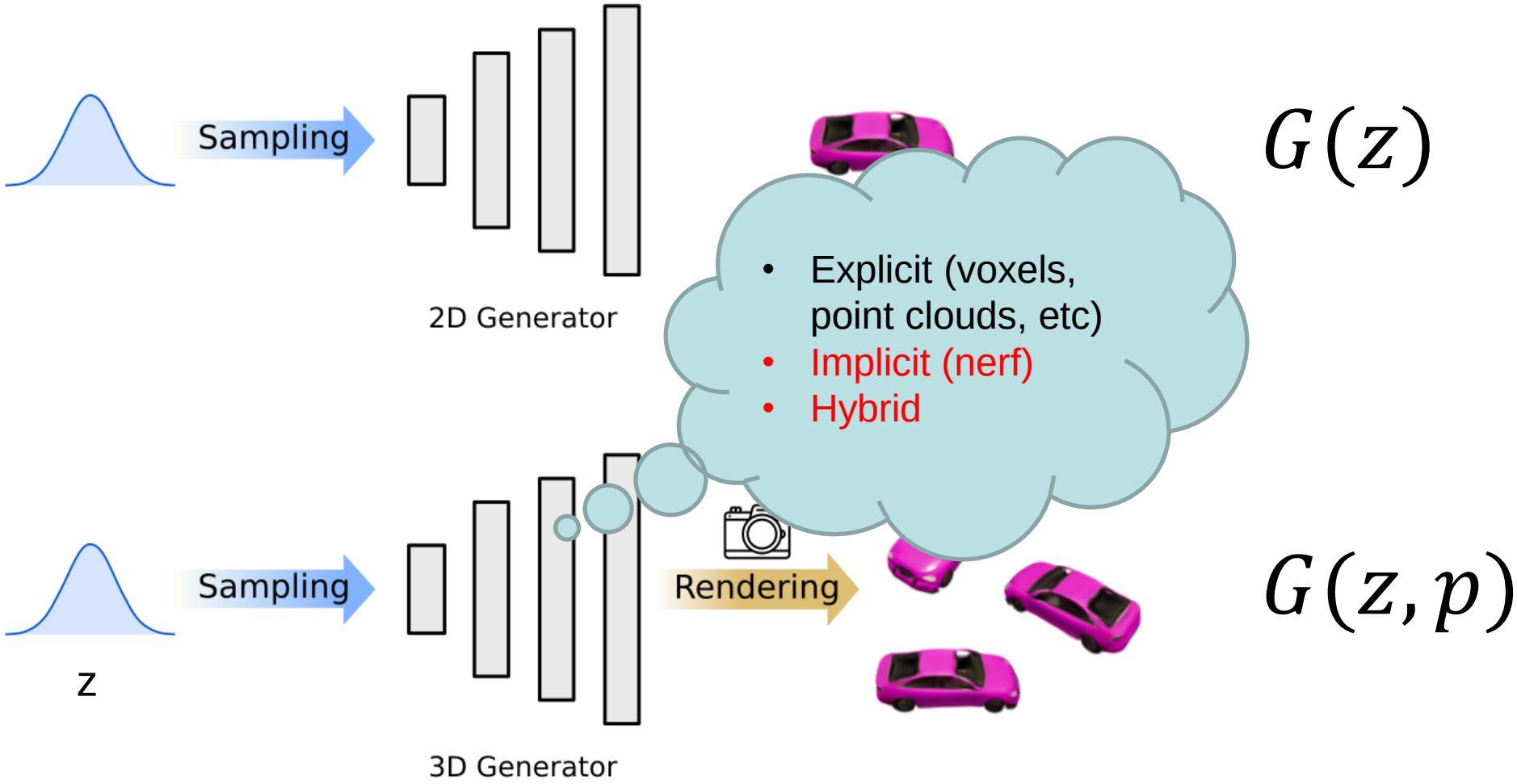
- Generative Models
  - Likelihood-based (VAEs, Flow, DDPM, Autoregressive models, etc)
  - Likelihood-free (GANs)

- Generative Adversarial Networks (GANs)



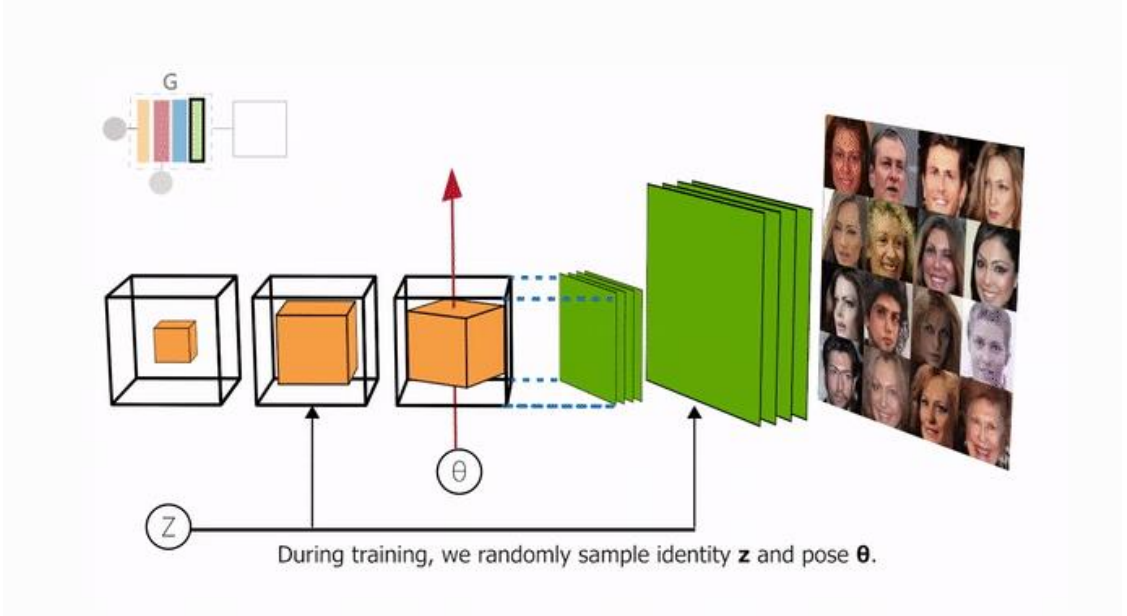
Motivation

- Making GANs 3D aware/consistent

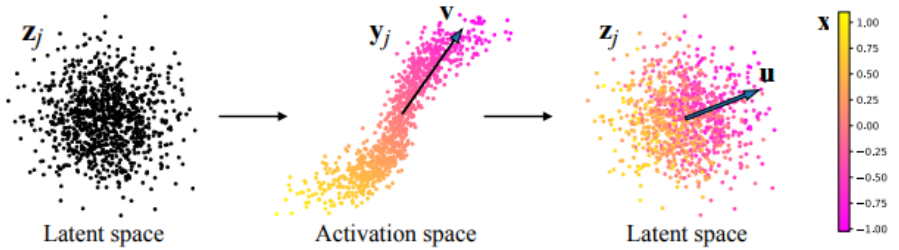


Existing Approaches

- Explicit representation of 3D GANs
  - HoloGAN
  - BlockGAN



- Manipulate the Latent space of 2D GANs
  - Latent/Style editing



Neural Radiance Fields (NeRF)

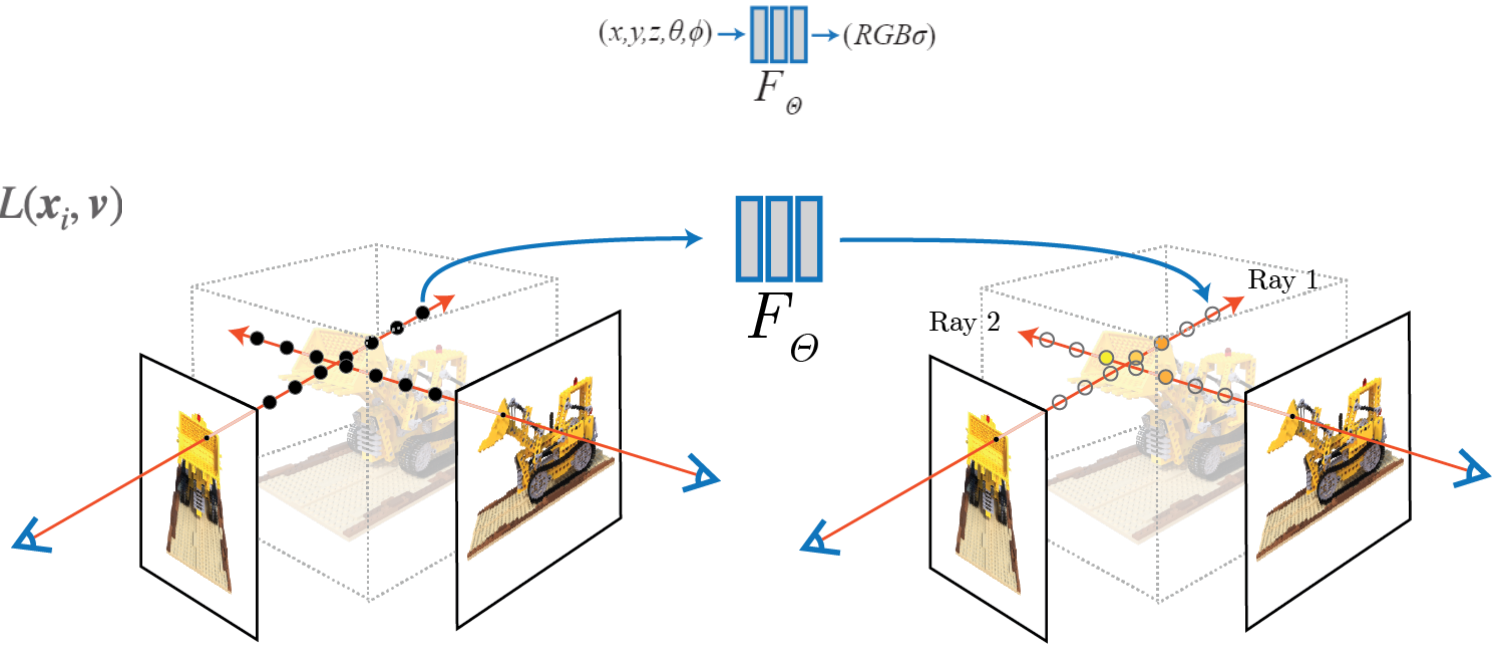
- Implicit fields
  - Input: point position, view direction
  - Output: RGB, density, features

- Volume rendering

$$\hat{I} = \mathbb{E}_{x\sim O} L(\boldsymbol{x}, \boldsymbol{v}) = \int_0^\infty \tau(t) L(\boldsymbol{x}(t), \boldsymbol{v}) dt \approx \sum_{i=1}^{n} \hat{\tau}_i L(\boldsymbol{x}_i, \boldsymbol{v})$$

$$\tau(t) = \frac{dO}{dt}(t)$$

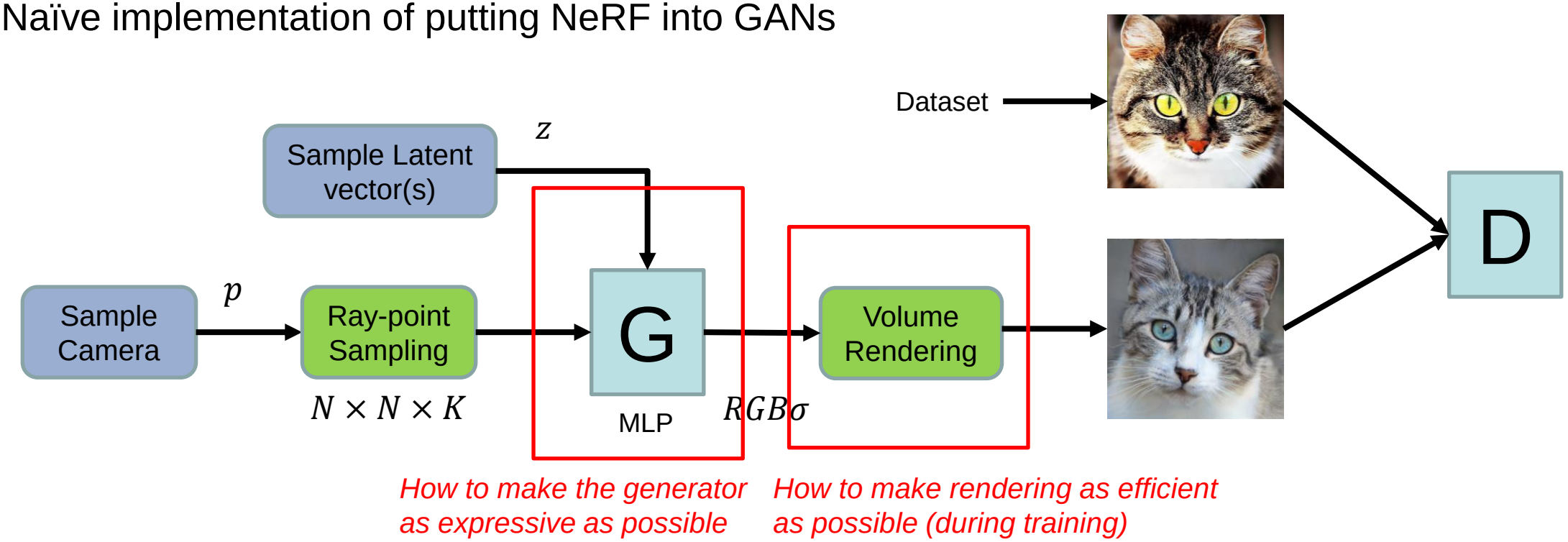
$$O(t) = 1 - \exp\Big(- \int_0^t \sigma(\boldsymbol{x}(s))\, ds \Big)$$



Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020, August). Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision* (pp. 405-421). Springer, Cham.

NeRF-GANs

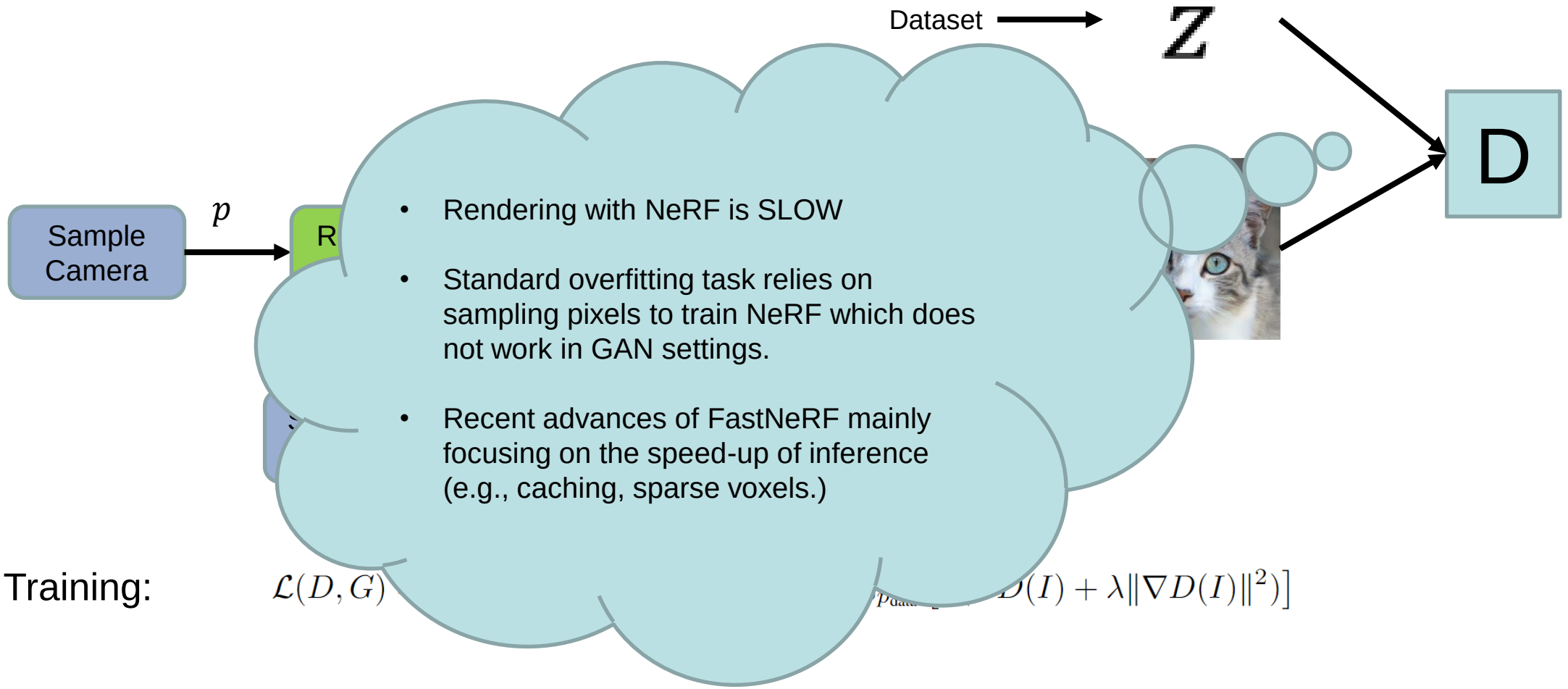
Naïve implementation of putting NeRF into GANs



Training: $\mathcal{L}(D,G) = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{Z},\boldsymbol{p}\sim\mathcal{P}}\left[f(D(G(\boldsymbol{z},\boldsymbol{p}))\right] + \mathbb{E}_{I\sim p_{\text{data}}}\left[f(-D(I) + \lambda\|\nabla D(I)\|^2)\right]$

NeRF-GANs

Naïve implementation of putting NeRF into GANs

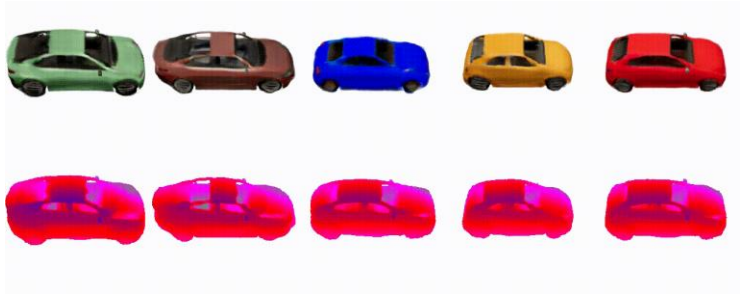


Training: $\mathcal{L}(D, G)$ $D(I) + \lambda \|\nabla D(I)\|^2)]$

Previous work: GRAF

- GRAF is AFAIK the first work combining NeRF in GAN framework
  - Simple MLP architecture, global Z concat with input position:
    - Not expressive enough to handle complex scenes
  - Sampling patches for discriminator to speed-up training
    - Worse performance on high-resolution images
    - Not really solve the rendering speed problem

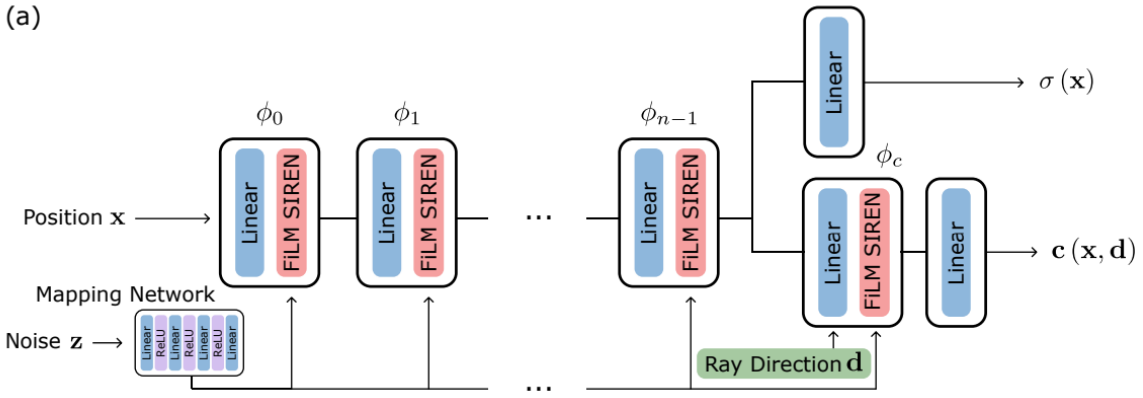
$s = 4$ $s = 2$ $s = 1$

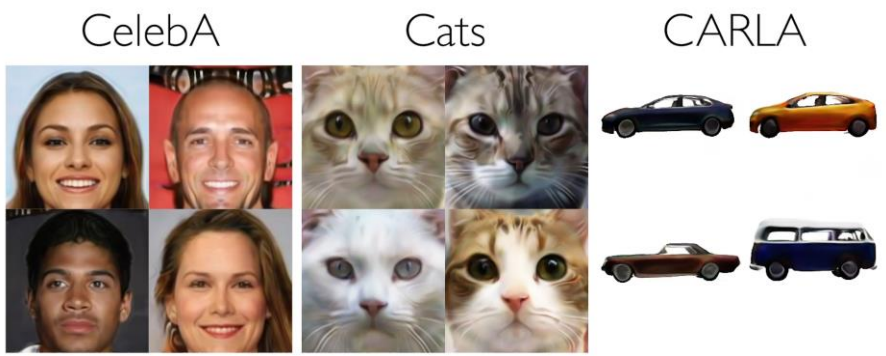


Schwarz, K., Liao, Y., Niemeyer, M., & Geiger, A. (2020). Graf: Generative radiance fields for 3d-aware image synthesis. arXiv preprint arXiv:2007.02442.

Previous work: Pi-GAN

- Pi-GAN achieves much better visual quality than GRAF
  - StyleGAN like mapping and synthesis network
  - SIREN activations
    - The model is trained at full resolution, so it is slow and can only work in low-resolution
    - The quality is still not good enough and far from 2D models
    - Inference is also slow

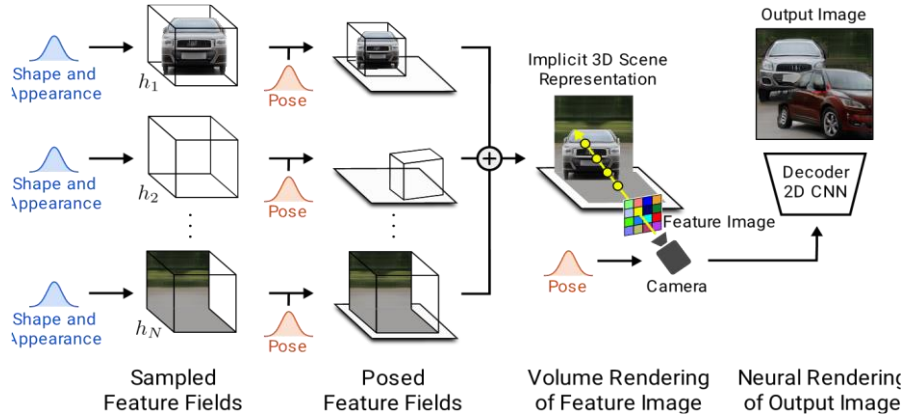


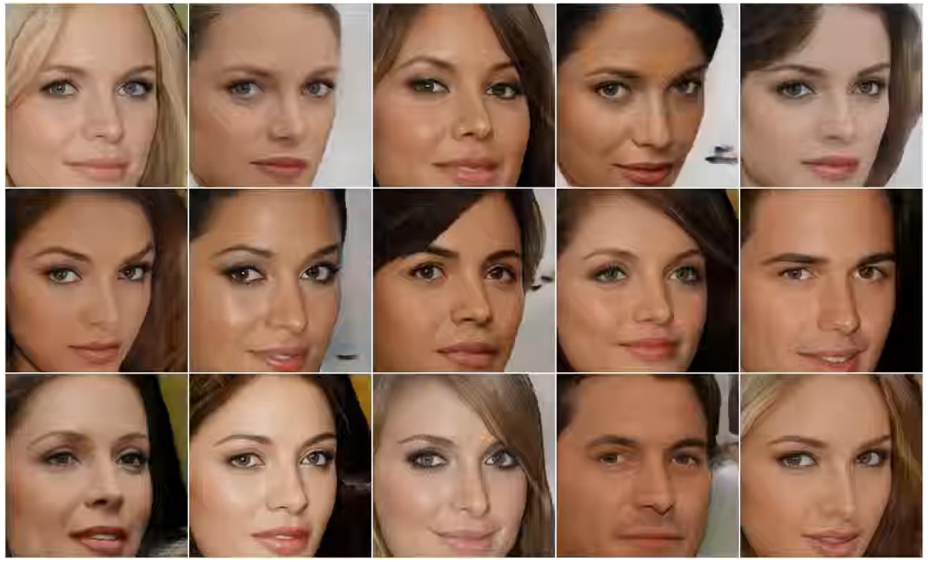


Chan, E. R., Monteiro, M., Kellnhofer, P., Wu, J., & Wetzstein, G. (2021). pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5799-5809).

Previous work: GiRAFFE

- GiRAFFE proposes a compositional NeRF with a 2D CNN decoder
  - Same as GRAF, z concat with position
  - NeRF works in 16x16 resolution
    - Light-weight, rendering is pretty fast
    - Architecture is too simple to handle complex scenes
    - 2D CNN causes serious inconsistency in the rendered outputs





Niemeyer, M., & Geiger, A. (2021). Giraffe: Representing scenes as compositional generative neural feature fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11453-11464).

Previous works

- GRAF, pi-GAN:
  - direct generative NeRF training

- Slow and only acceptable at low-resolution
- Poor visual quality

- GiRAFFE
  - NeRF + CNN upsampler

- Poor multi-view consistency

StyleNeRF: A Style-based 3D-Aware Generator for High-resolution Image Synthesis

Jiatao Gu, Lingjie Liu, Peng Wang, Christian Theobalt
Meta AI, Max Planck Institute, University of Hong Kong
Arxiv 2021

Website: https://jiataogu.me/style_nerf/

Comparison with Existing works

Comparisons: FFHQ Dataset (256x256)

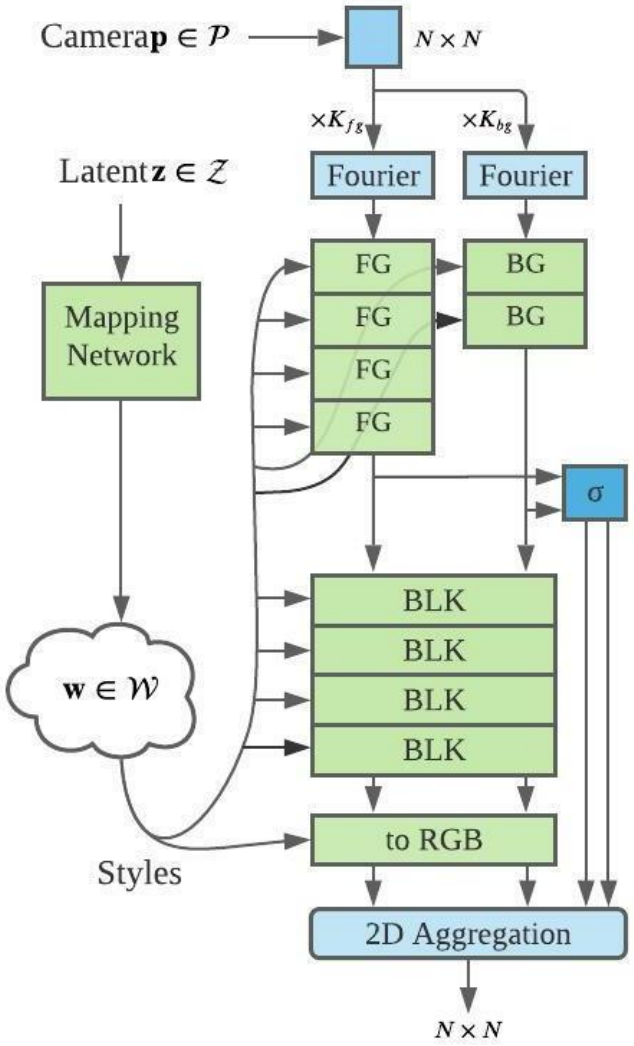
HoloGAN

Goal

- We propose to address the above issues simultaneously:
  - High-resolution
  - Efficient
  - Multi-view consistent

- We achieve this by incorporating a StyleGAN2-like architecture with approximation, while carefully designing strategies to maintain consistency

Architecture



- The basic architecture is like pi-GAN, while we use LeakyReLU and style modulation to build the model.
- We follow NeRF++, use a separate nerf to model background*
- We split the network of predicting density and color into two parts, while each part is style-modulated.

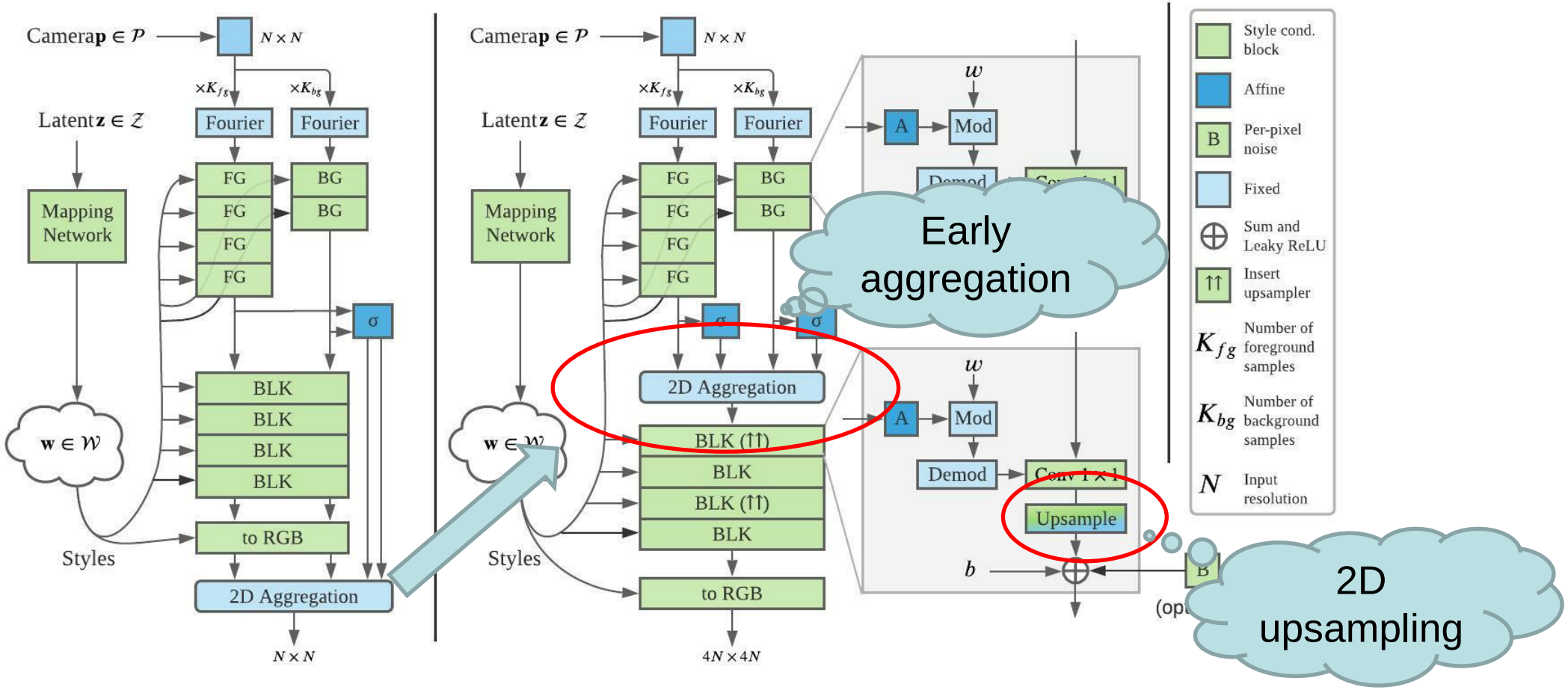
$$\boldsymbol{\phi}_{\boldsymbol{w}}^{n}(\boldsymbol{x}) = g_{\boldsymbol{w}}^{n} \circ g_{\boldsymbol{w}}^{n-1} \circ \ldots \circ g_{\boldsymbol{w}}^{1} \circ \zeta(\boldsymbol{x})$$

$$\boldsymbol{c}_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{d}) = h_c \circ [\boldsymbol{\phi}_{\boldsymbol{w}}^{n_c}(\boldsymbol{x}), \zeta(\boldsymbol{d})]$$

$$\sigma_{\boldsymbol{w}}(\boldsymbol{x}) = h_\sigma \circ \boldsymbol{\phi}_{\boldsymbol{w}}^{n_\sigma}(\boldsymbol{x})$$

* Recently we also tried using single nerf with depth warping similar as https://arxiv.org/abs/2111.12077
It also works quite well modeling distant background.

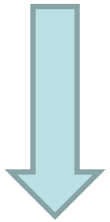
Architecture



Architecture

- Approximated Volume Rendering

$$I_{\boldsymbol{w}}^{\text{NeRF}}(\boldsymbol{r}) = \int_0^{\infty} p_{\boldsymbol{w}}(t)\boldsymbol{c}_{\boldsymbol{w}}(\boldsymbol{r}(t), \boldsymbol{d})dt, \quad \text{where} \quad p_{\boldsymbol{w}}(t) = \exp\left(-\int_0^{t} \sigma_{\boldsymbol{w}}(\boldsymbol{r}(s))ds\right) \cdot \sigma_{\boldsymbol{w}}(\boldsymbol{r}(t))$$



$$I_{\boldsymbol{w}}^{\text{Approx}}(\boldsymbol{r}) = \int_0^{\infty} p_{\boldsymbol{w}}(t) \cdot h_c \circ \left[\phi_{\boldsymbol{w}}^{n_c}(\boldsymbol{r}(t)), \zeta\left(\boldsymbol{d}\right)\right] dt \approx h_c \circ \left[\phi_{\boldsymbol{w}}^{n_c, n_\sigma}\left(\mathcal{A}(\boldsymbol{r})\right), \zeta\left(\boldsymbol{d}\right)\right]$$

Early aggregation

$$\phi_{\boldsymbol{w}}^{n, n_\sigma}\left(\mathcal{A}(R_H)\right) \approx \texttt{Upsample}\left(\phi_{\boldsymbol{w}}^{n, n_\sigma}\left(\mathcal{A}(R_L)\right)\right)$$

2D upsampling

Maintain 3D consistency

- Remove view direction input
  - We found that view direction will break the consistency and did not contribute to much quality (our dataset is single image)

- Up-sampler design

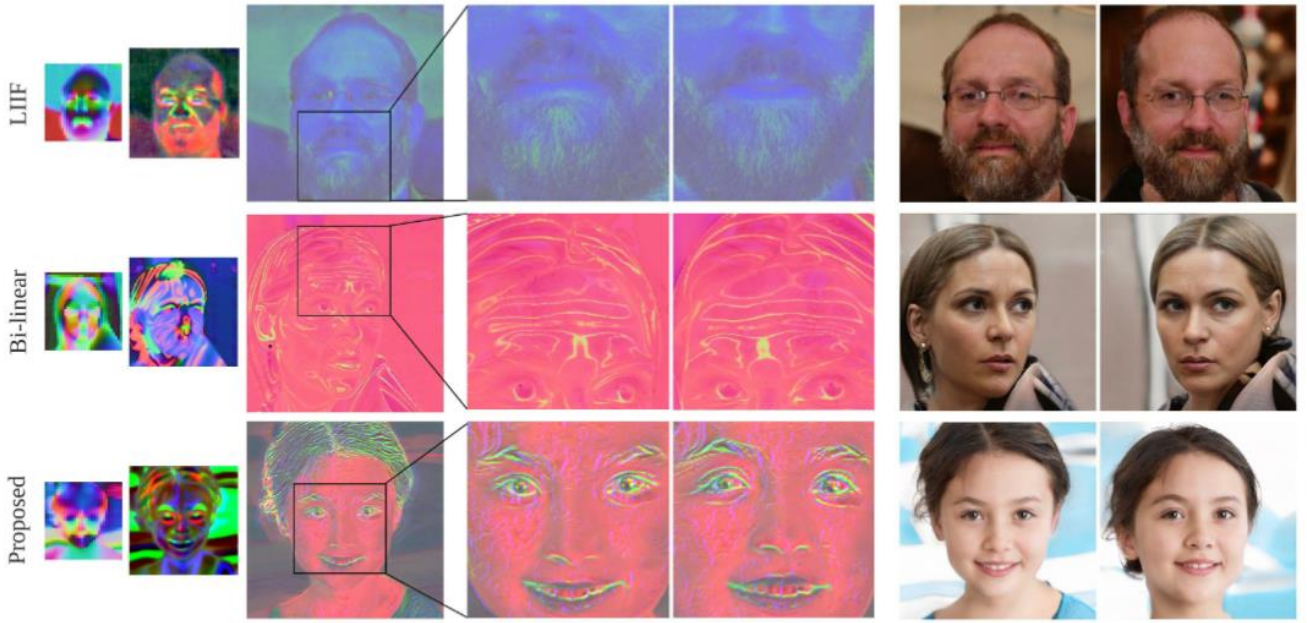
$$\texttt{Upsample}(X) = \texttt{Conv2d}\left(\texttt{Pixelshuffle}\left(\texttt{Repeat}(X, 4) + \psi_\theta(X), 2\right), K\right)$$

- NeRF-path regularization

$$\mathcal{L}_{\text{NeRF-path}} = \frac{1}{|S|} \sum\nolimits_{(i,j)\in S} \left(I_{\boldsymbol{w}}^{\text{Approx}}(R_{\text{in}})[i,j] - I_{\boldsymbol{w}}^{\text{NeRF}}(R_{\text{out}}[i,j])\right)^2$$

StyleNeRF

- Up-sampler: we have tested many ways
  - Filter-based (bilinear interpolation, FIR filters, etc) + MLP (1x1 Conv) will cause “bubble shape” artifacts
  - Learning-based (transposed conv, pixelshuffle, LIIF) will easily cause texture sticking artifacts
  - We combine these two methods



Meta AI

Ablation: Different Upsampling Operators

LIIF: Having the "texture sticking" artifacts

Meta AI

Ablation: Different Upsampling Operators

Bilinear: Having the "bubble-shape" artifacts

Meta AI

Ablation: Different Upsampling Operators

Our proposed operator: Highly preserving 3D consistency while getting rid of bubble-shape artifacts

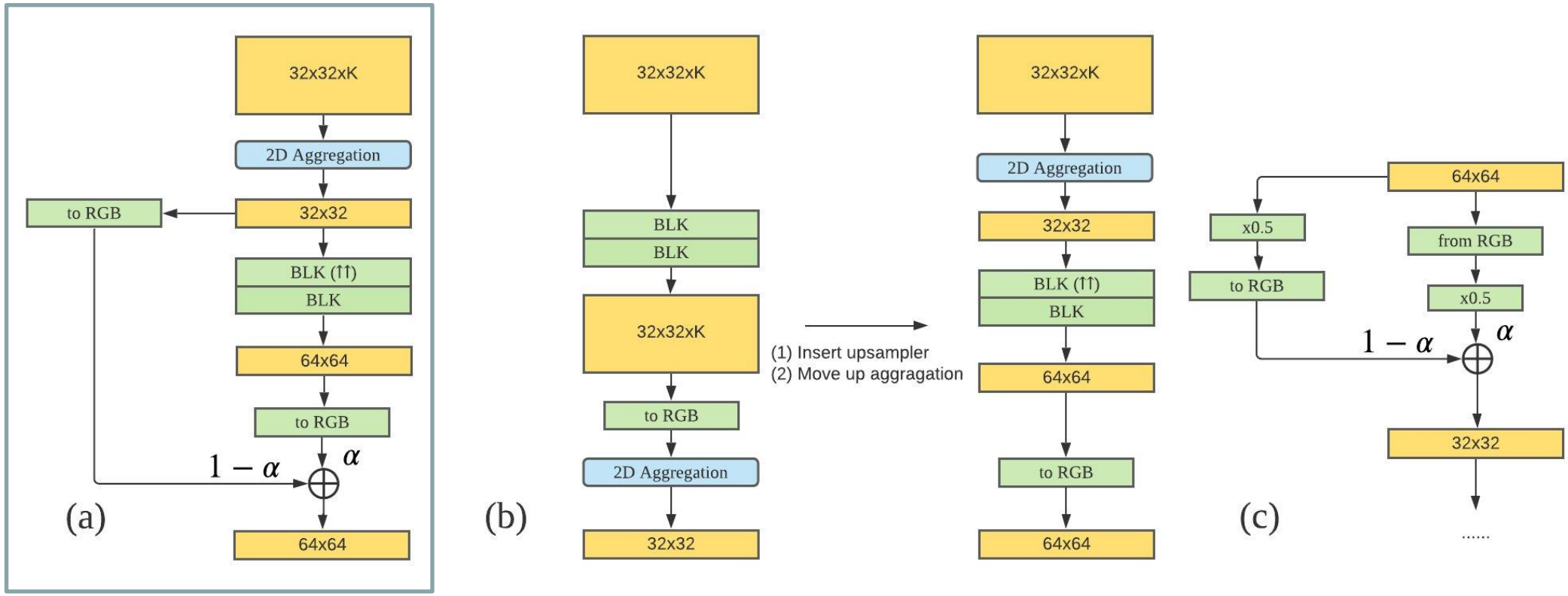
StyleNeRF

- Noise Injection
  - S
    i
    r
  - H
    i
    g

Ablation: Noise Injection

StyleNeRF

- Progressive Training



Results (rendering)

Our Results: FFHQ Dataset (1024x1024)

This is the first time that a generative model can synthesize high-resolution images from novel views while preserving high 3D consistency

Results

- V.s. Existing works

Models	FFHQ 256^2		AFHQ 256^2		CompCars 256^2		Rendering time (ms / image)				
	FID	KID	FID	KID	FID	KID	64	128	256	512	1024
2D GAN	4	1.1	9	2.3	3	1.6	-	-	46	51	53
HoloGAN	75	68.0	78	59.4	48	39.6	213	215	222	-	-
GRAF	71	57.2	121	83.8	101	86.7	61	246	990	3852	15475
π-GAN	85	90.0	47	29.3	295	328.9	58	198	766	3063	12310
GIRAFFE	35	23.7	31	13.9	32	23.8	8	-	9	-	-
Ours	8	3.7	14	3.5	8	4.3	-	-	65	74	98

- High resolution

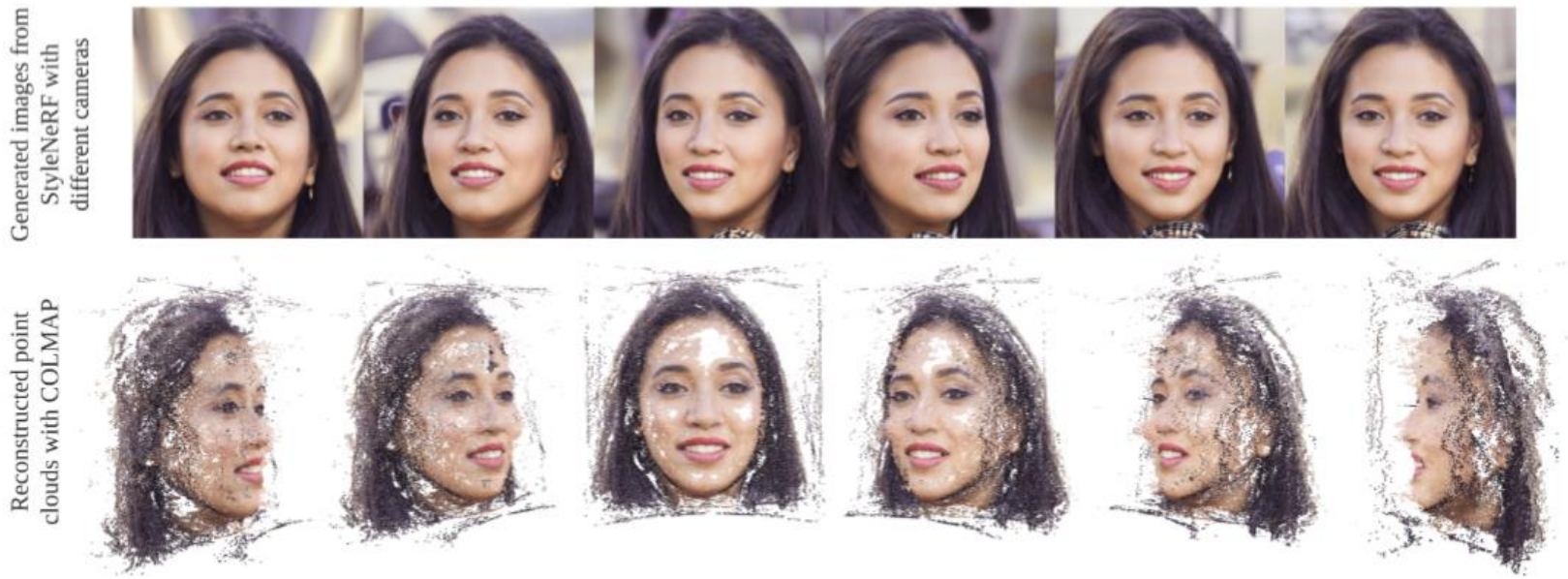
Models	FFHQ 512^2		AFHQ 512^2		MetFace 512^2		FFHQ 1024^2	
	FID	KID	FID	KID	FID	KID	FID	KID
2D GAN	3.1	0.7	8.6	1.7	18.9	2.7	2.7	0.5
Ours	7.8	2.2	13.2	3.6	20.4	3.3	8.1	2.4

Results

- Consistency evaluation (quantitively/qualitatively)

Table 3: PSNR, SSIM and LPIPS scores between the images synthesized by 3D-aware generative models, and the corresponding pseudo targets generated by IBRNet (Wang et al., 2021).

Models	FFHQ 256^2			AFHQ 256^2		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
π-GAN	29.5	0.92	0.04	28.5	0.86	0.12
GIRAFFE	25.8	0.81	0.13	26.2	0.75	0.25
Ours	29.0	0.89	0.08	26.8	0.80	0.13



Results (interpolation)

Style Interpolation

Our synthesized results (512x512)

Results (style mixing/inversion)

Results

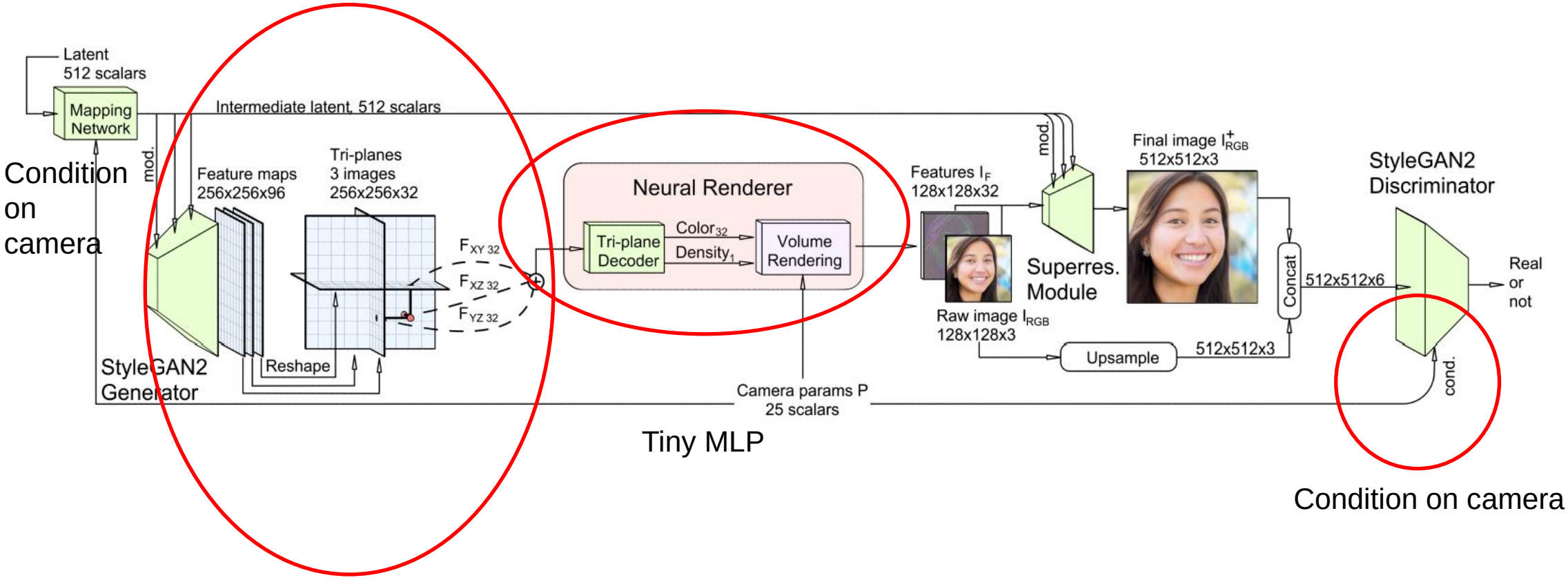
- Interactive Demo

http://127.0.0.1:7111/

EMERGING RECENT WORKS

Concurrent / Recent works: EG3D

- https://matthew-a-chan.github.io/EG3D/



Condition on camera

Tiny MLP

Condition on camera

Tri-plane representations

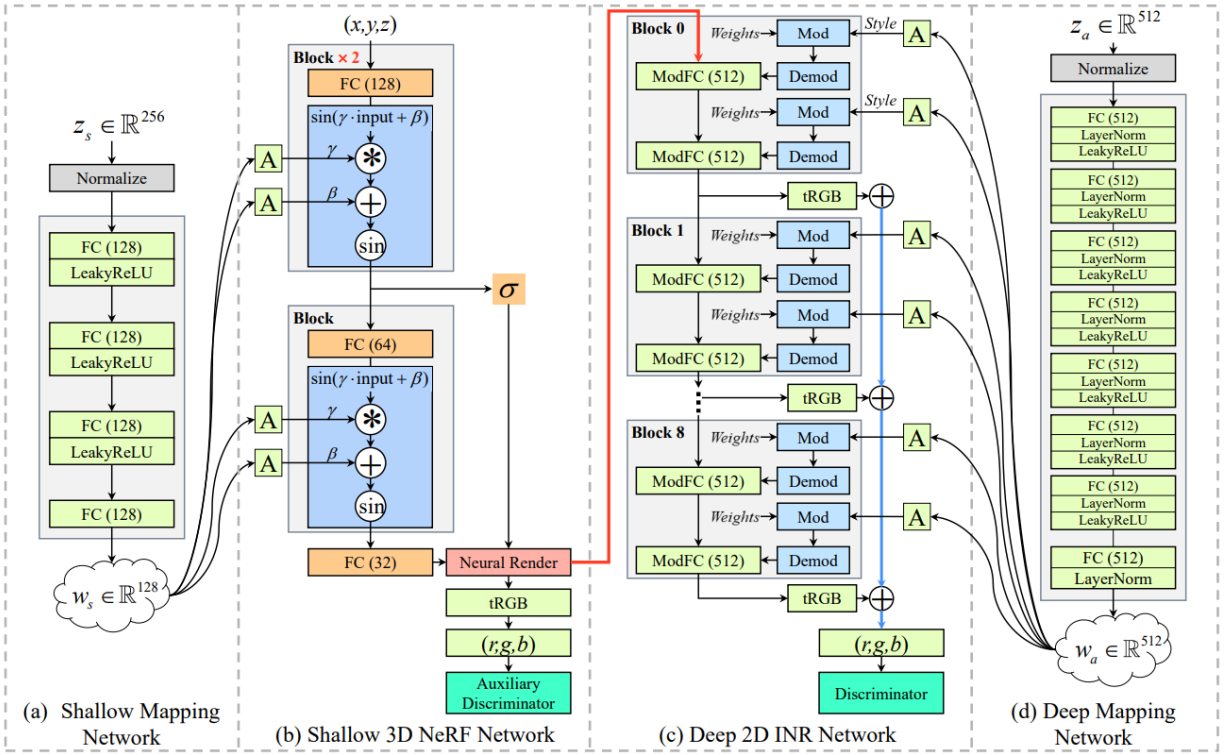
Concurrent / Recent works: EG3D

- EG3D achieves much better visual quality and Multiview consistency compared to our works
  - In our view, it is mainly because it renders under a resolution of 128x128

- However, this method needs camera input.

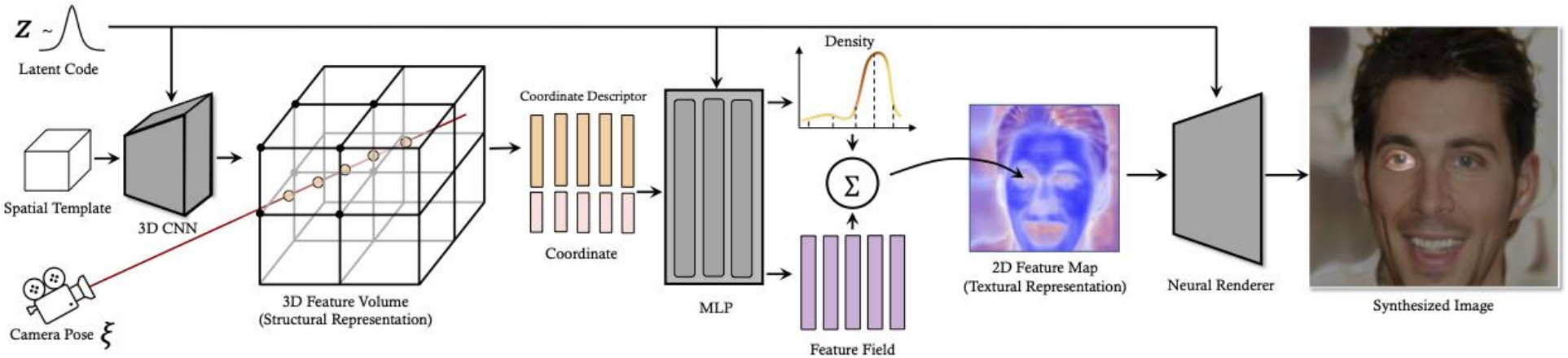
Concurrent / Recent works: CIPS-3D

- https://github.com/PeterouZh/CIPS-3D
- Similar to StyleNeRF, but without using up-sampling.
  - Tiny NeRF + Deep 2D CNNs (pi-GAN + CIPS)



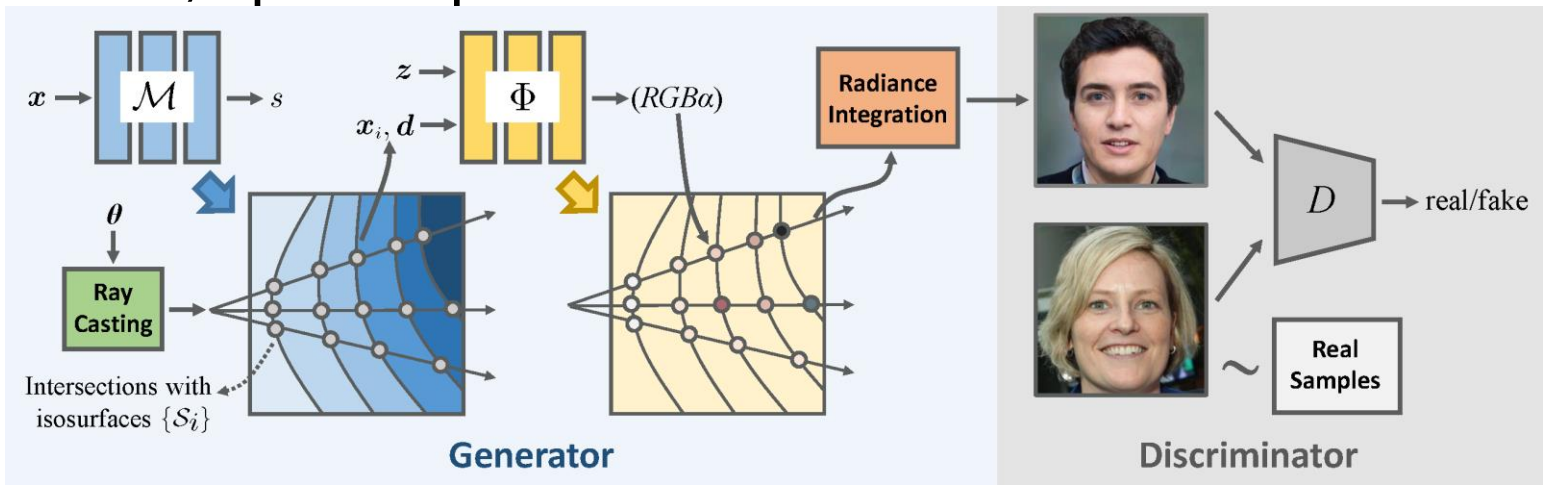
Concurrent / Recent works: VolumeGAN

- https://genforce.github.io/volumegan/
- Similar to StyleNeRF and EG3D, while using a volume (3D CNNs) as the explicit representation.



Concurrent / Recent works: GRAM

- https://yudeng.github.io/GRAM/
- Similar to pi-GAN, pure NeRF model
- Using small network to generate manifold to guild the sampling of big network. However, speed-up is unknown.



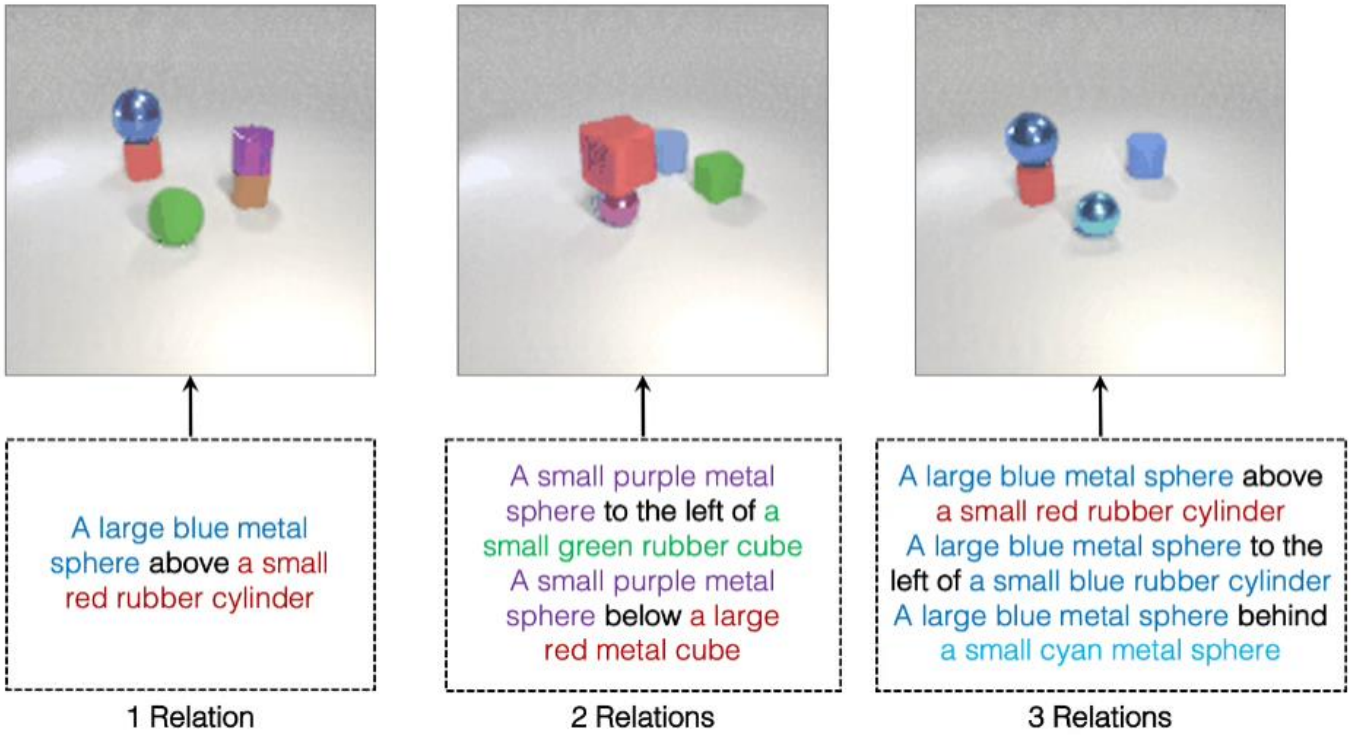
Future Directions

- 3D-aware generative models on more diverse datasets

Leveraging existing 3D datasets

Future Directions

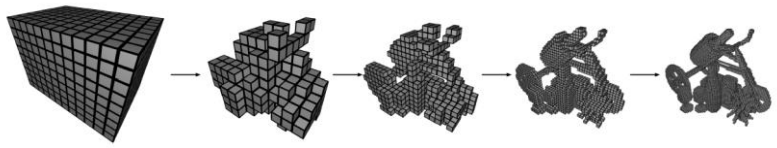
- Text-to-3D generation
- Language learning via 3D generation



More work on Implicit Representations

- Along the research line of implicit representation, we have also explored aspects across different applications:

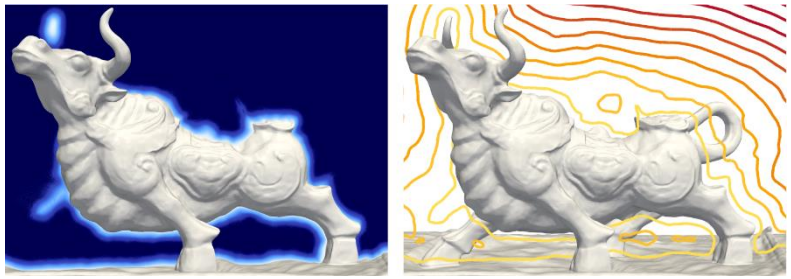
Efficient/Editable Rendering



Neural Sparse Voxel Field
Liu et al. NeurIPS 2020 (spotlight)

Exploring efficient and editable NeRF rendering with sparse voxel octree.

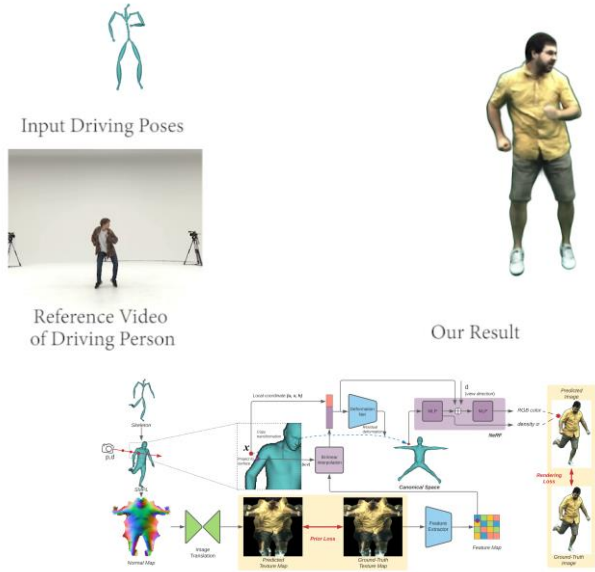
Geometry Modeling



Volume Rendering of Neural Implicit Surfaces
Yariv et al. NeurIPS 2021 (oral)

We propose VolSDF to learn surfaces using NeRF-like volume rendering

Human NeRF



Neural Actor: Neural Free-view Synthesis of Human Actors with Pose Control
Liu et al. SIGGRAPH ASIA 2021

We propose Neural Actor for high quality synthesis of humans using geometry guidance and latent texture map.

Meta AI

Thank you! Q & A