Generative Neural Scene Representations for 3D-Aware Image Synthesis

Andreas Geiger

Autonomous Vision Group
University of Tübingen and MPI for Intelligent Systems

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

MAX-PLANCK-GESELLSCHAFT

ellis
European Laboratory for Learning and Intelligent Systems

Covered Papers

GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis
Katja Schwarz and Yiyi Liao and Michael Niemeyer and Andreas Geiger
NeurIPS 2020

GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields
Michael Niemeyer, Andreas Geiger
CVPR 2021

CAMPARI: Camera-Aware Decomposed Generative Neural Radiance Fields
Michael Niemeyer, Andreas Geiger
3DV 2021

Collaborators

Michael Niemeyer

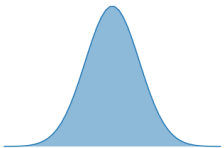
Katja Schwarz

Yiyi Liao

Generative models are fantastic!

Generative Models

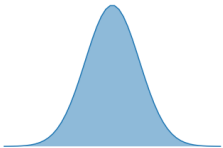
Sample a latent code from the prior distribution.



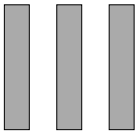
Latent Code

Generative Models

Pass latent code to trained generator G_θ.



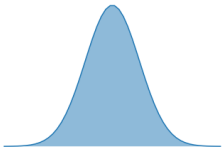
Latent Code



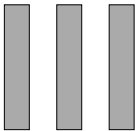
Generator G_θ

Generative Models

The generator outputs a synthesized image.



Latent Code



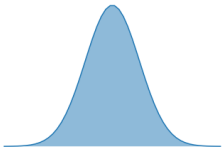
Generator G_θ

Generated Image*

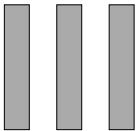
* The generated images are samples from StyleGAN2.

Generative Models

Sample more latent codes to get different generated images.



Latent Code



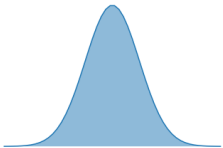
Generator G_θ

Generated Image*

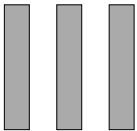
* The generated images are samples from StyleGAN2.

Generative Models

Sample more latent codes to get different generated images.



Latent Code



Generator G_θ

Generated Image*

* The generated images are samples from StyleGAN2.

Is generating photorealistic images all we need?

Generative Models

For many applications, we require **control over the image generation process:**

Generative Models

For many applications, we require **control over the image generation process:**

Animation Movies

Video Source: Disney's Toy Story 4 Trailer

Generative Models

For many applications, we require **control over the image generation process:**

Video Games

Video Source: Gran Turismo 7 Trailer

Generative Models

For many applications, we require **control over the image generation process:**

Virtual Reality

Video Source: Oculus Rift Trailer

Generative Models

Goal: A generative model for **3D-aware image synthesis** which allows us to:

Generative Models

Goal: A generative model for **3D-aware image synthesis** which allows us to:

- Generate photorealistic images

Generative Models

Goal: A generative model for **3D-aware image synthesis** which allows us to:

- Generate photorealistic images
- Control individual objects wrt. their appearance, pose and size in 3D

Generative Models

Goal: A generative model for **3D-aware image synthesis** which allows us to:

- Generate photorealistic images
- Control individual objects wrt. their appearance, pose and size in 3D
- Control the camera viewpoint in 3D

Generative Models

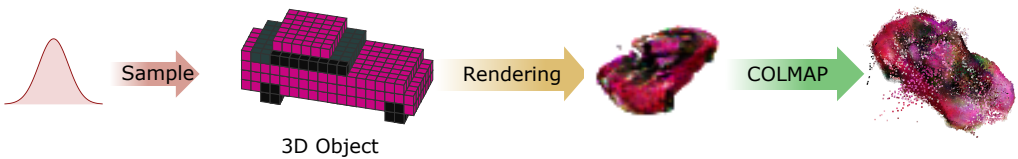
Goal: A generative model for **3D-aware image synthesis** which allows us to:

- Generate photorealistic images
- Control individual objects wrt. their appearance, pose and size in 3D
- Control the camera viewpoint in 3D
- Train from unstructured and unposed images

What representation should we use for
3D-aware image synthesis?

3D Representations

Voxel-based 3D Shape with Volumetric Rendering

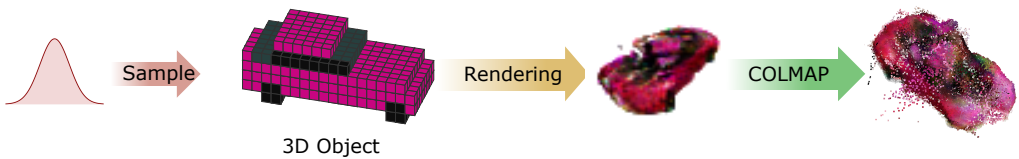


PlatonicGAN [Henzler et al., ICCV 2019]

Schwarz, Liao, Niemeyer, Geiger: GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. NeurIPS, 2020.

3D Representations

Voxel-based 3D Shape with Volumetric Rendering

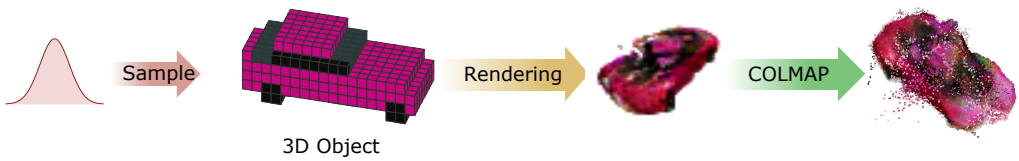


PlatonicGAN [Henzler et al., ICCV 2019]

+ Multi-view consistent

3D Representations

Voxel-based 3D Shape with Volumetric Rendering



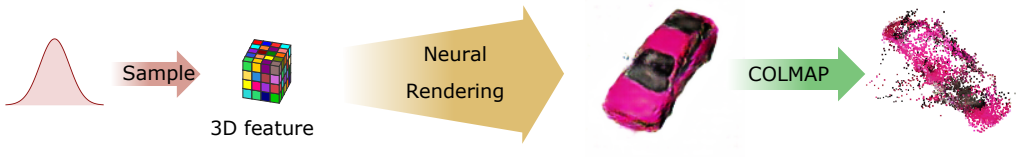
PlatonicGAN [Henzler et al., ICCV 2019]

- \+ Multi-view consistent
- − Low image fidelity, high memory consumption

Schwarz, Liao, Niemeyer, Geiger: GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. NeurIPS, 2020.

3D Representations

Voxel-based 3D Latent Feature with Learnable Projection

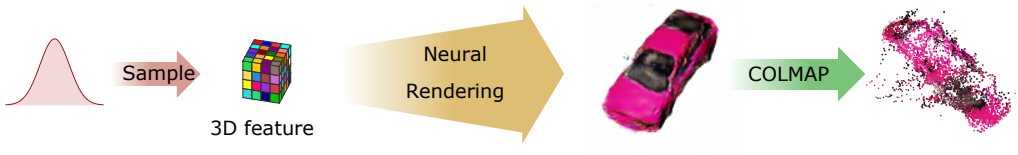


HoloGAN [Nguyen-Phuoc et al., ICCV 2019]

Schwarz, Liao, Niemeyer, Geiger: GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. NeurIPS, 2020.

3D Representations

Voxel-based 3D Latent Feature with Learnable Projection

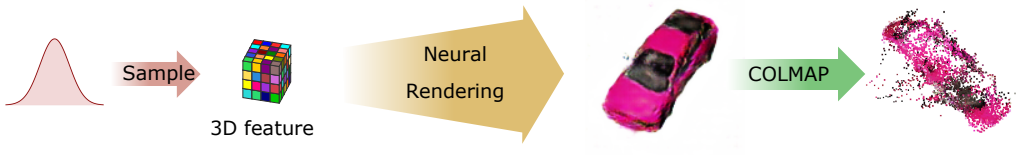


HoloGAN [Nguyen-Phuoc et al., ICCV 2019]

+ High image fidelity

3D Representations

Voxel-based 3D Latent Feature with Learnable Projection



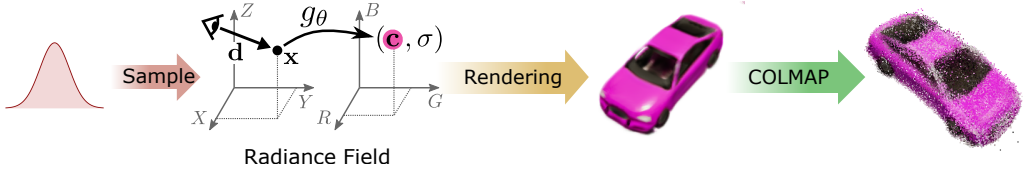
HoloGAN [Nguyen-Phuoc et al., ICCV 2019]

- \+ High image fidelity
- − Object shape and identity vary with viewpoint due to learnable projection

Schwarz, Liao, Niemeyer, Geiger: GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. NeurIPS, 2020.

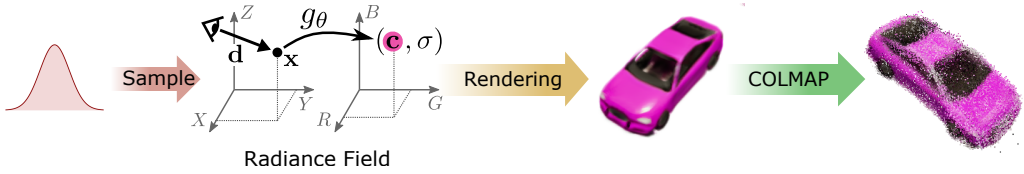
3D Representations

Generative Radiance Fields



3D Representations

Generative Radiance Fields

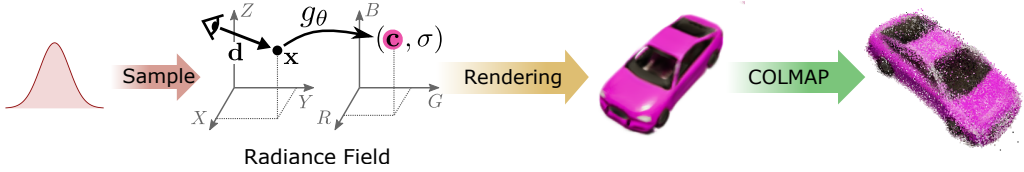


+ Continuous representation, multi-view consistent

Schwarz, Liao, Niemeyer, Geiger: GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. NeurIPS, 2020.

3D Representations

Generative Radiance Fields



- Continuous representation, multi-view consistent
- High image fidelity, low memory consumption

Schwarz, Liao, Niemeyer, Geiger: GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. NeurIPS, 2020.

GRAF: Generative Radiance Fields

Sample camera matrix $\mathbf{K}$, camera pose $\boldsymbol{\xi} \sim p_\xi$, and patch sampling pattern $\boldsymbol{\nu} \sim p_\nu$.

$\mathbf{K}$

$\boldsymbol{\xi} \sim p_\xi$

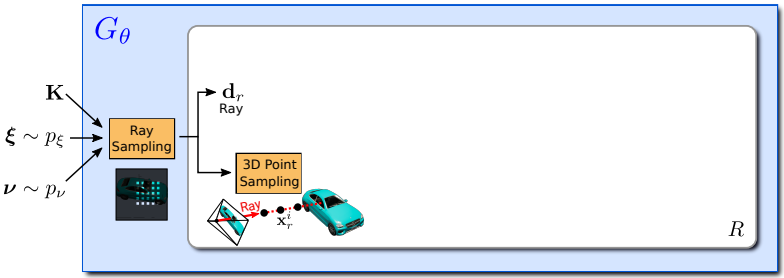
$\boldsymbol{\nu} \sim p_\nu$

GRAF: Generative Radiance Fields

Pass $\mathbf{K}$, $\boldsymbol{\xi}$, and $\boldsymbol{\nu}$ to generator G_θ which samples corresponding pixels / rays.

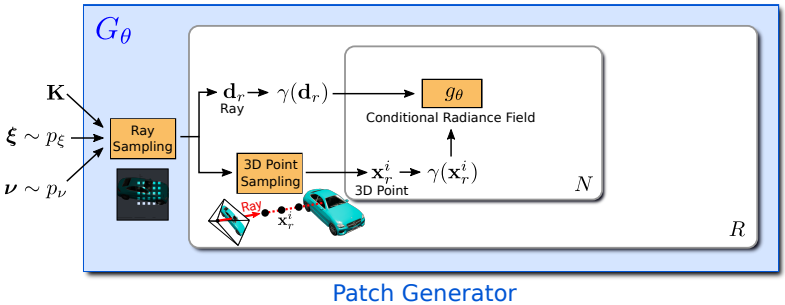
GRAF: Generative Radiance Fields

For each ray, get viewing direction $\mathbf{d}_r$ and sample 3D points $\mathbf{x}_r^i$ along ray.



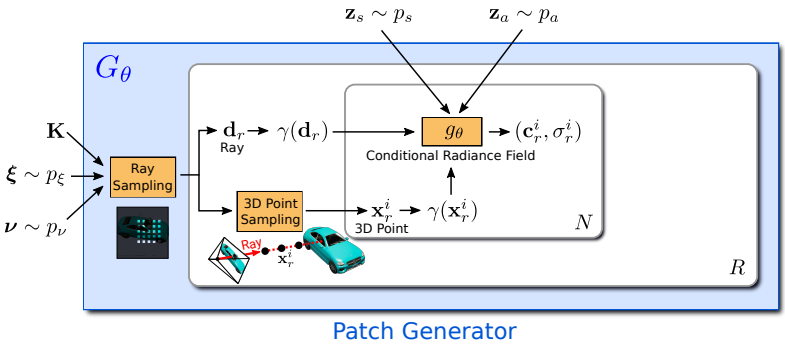
GRAF: Generative Radiance Fields

Push $\mathbf{d}_r$ and $\mathbf{x}_r^i$ through positional encoding γ as input to conditional radiance field g_θ.



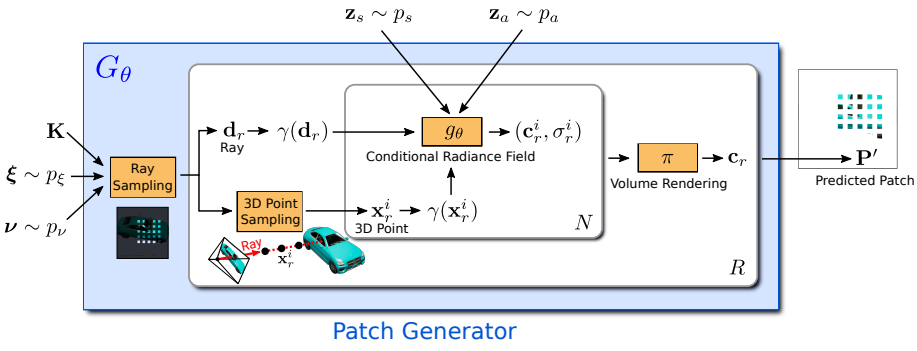
GRAF: Generative Radiance Fields

Sample latent shape and appearance codes $\mathbf{z}_s, \mathbf{z}_a$ and pass them as condition to g_θ.



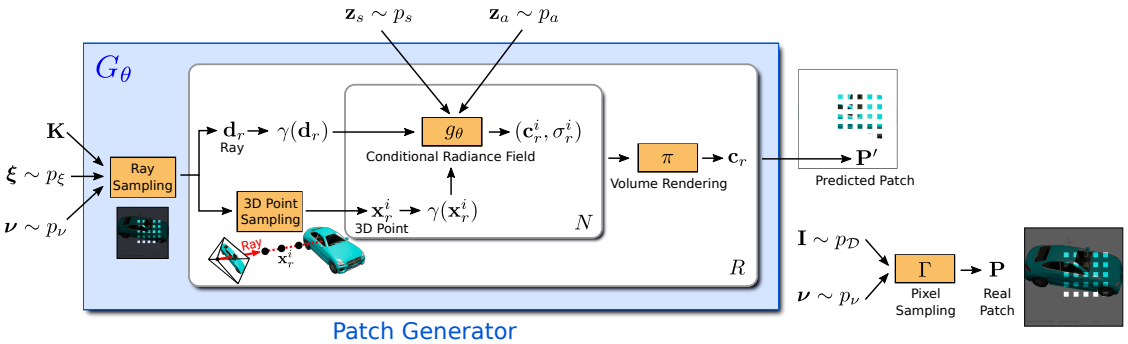
GRAF: Generative Radiance Fields

Perform volume-rendering for each ray, resulting in predicted patch $\mathbf{P}'$.



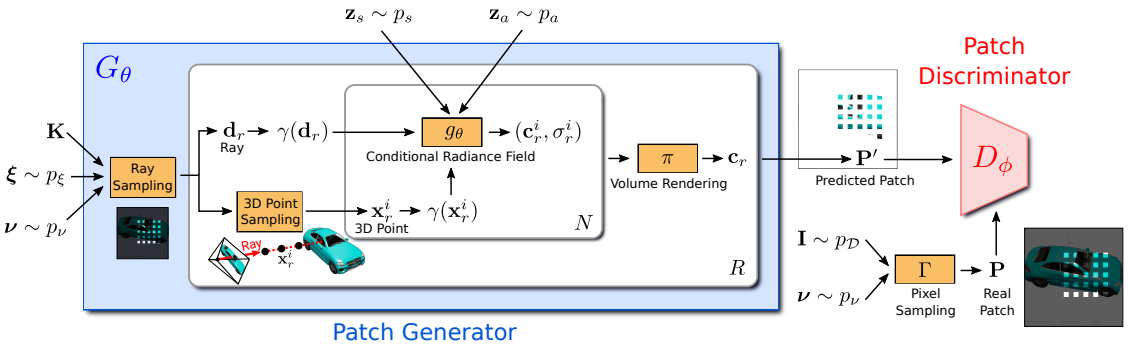
GRAF: Generative Radiance Fields

Sample patch $\mathbf{P}$ from real image $\mathbf{I}$ drawn from the data distribution $p_{\mathcal{D}}$ with pattern $\boldsymbol{\nu}$.

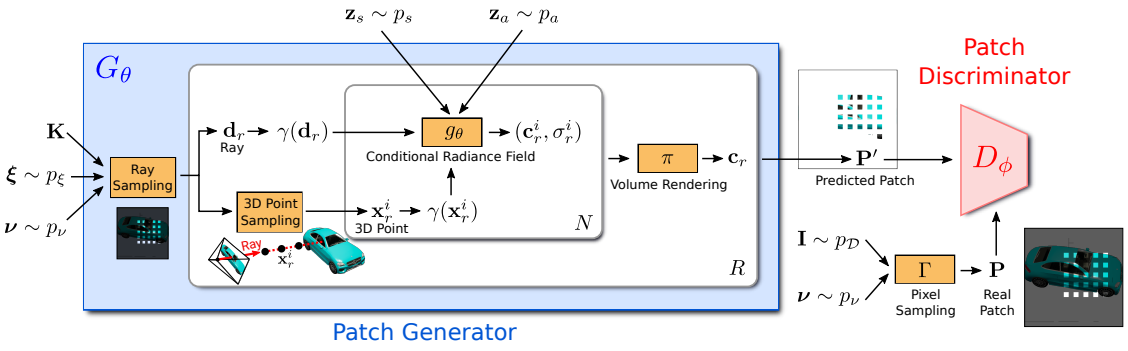


GRAF: Generative Radiance Fields

Pass fake and real patch $\mathbf{P}', \mathbf{P}$ to discriminator D_ϕ and train with adversarial loss.

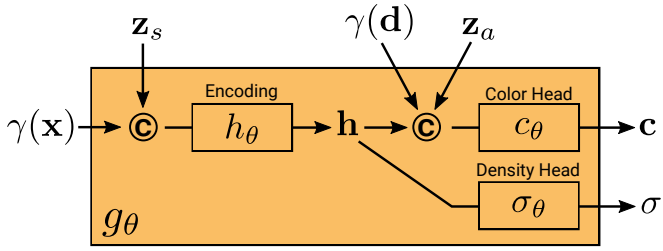


GRAF: Generative Radiance Fields



- Key idea: Generator/discriminator for **image patches** of size 32×32 pixels
- Patches sampled at **random locations** and with **random scales** (using dilation)

Conditional Radiance Fields



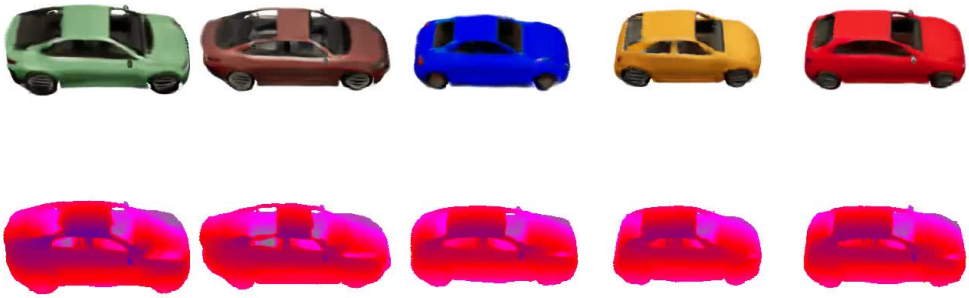
Conditional Radiance Field Architecture:

- Fully-connected MLP g_θ with ReLU activations
- Appearance code $\mathbf{z}_a$ and view direction influence only the output color
- Inductive biases allow the model to disentangle shape and appearance

Results

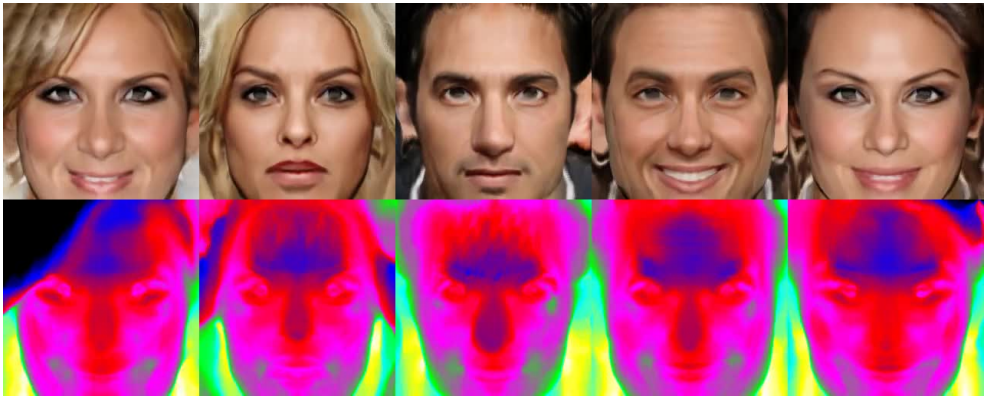
Generative Radiance Fields

Results on synthetic Carla dataset at 256^2 pixels:



Generative Radiance Fields

Results on real CelebA-HQ dataset at 256^2 pixels:



Schwarz, Liao, Niemeyer, Geiger: GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. NeurIPS, 2020.

How can we scale this idea to
more complex, multi-object scenes?

GIRAFFE: Compositional Generative Neural Feature Fields

GRAF:

- Incorporate a **3D representation** into the generative model

GIRAFFE: Compositional Generative Neural Feature Fields

GRAF:

- Incorporate a **3D representation** into the generative model

GIRAFFE:

- Incorporate a **compositional 3D scene representation** into the generative model

GIRAFFE: Compositional Generative Neural Feature Fields

GRAF:

- Incorporate a **3D representation** into the generative model

GIRAFFE:

- Incorporate a **compositional 3D scene representation** into the generative model
- Incorporate a **neural renderer** to yield fast and high-quality inference

GIRAFFE

Sample N shape and appearance codes.

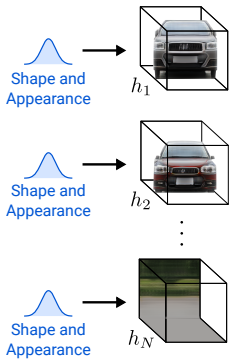
Shape and Appearance

Shape and Appearance

Shape and Appearance

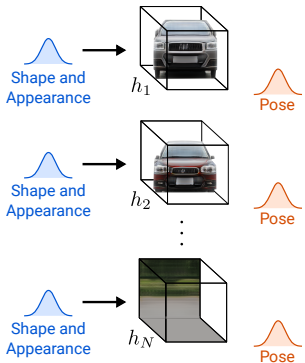
GIRAFFE

Get N **feature fields.** Note: We show features in RGB color for clarity.



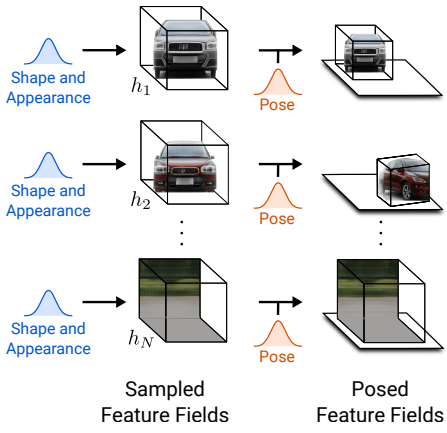
GIRAFFE

Sample size and pose for each feature field.



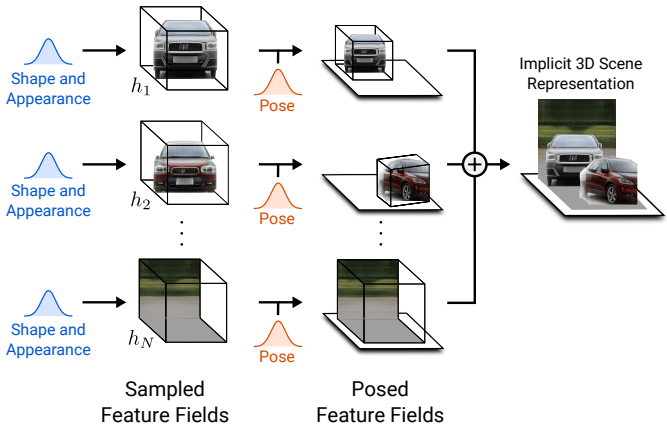
GIRAFFE

Get posed feature fields.



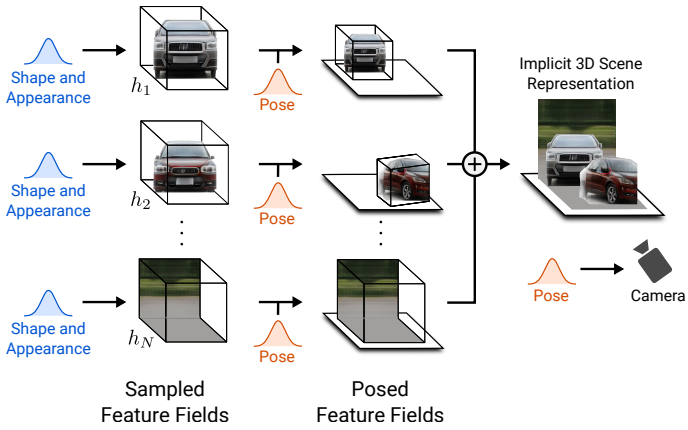
GIRAFFE

Composite all feature feature fields into a single 3D scene representation.



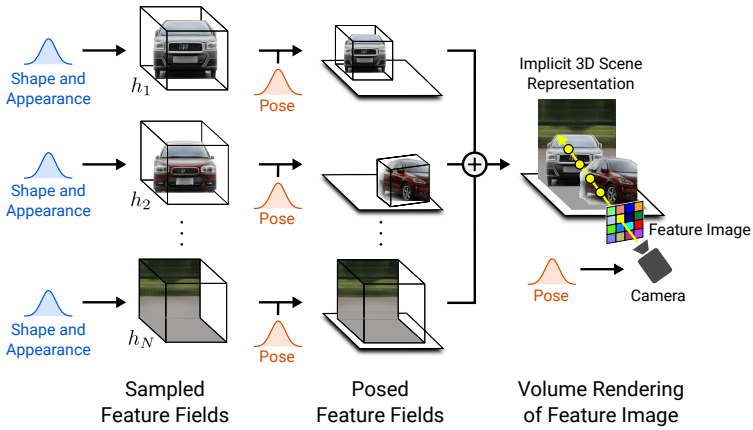
GIRAFFE

Sample a camera pose.



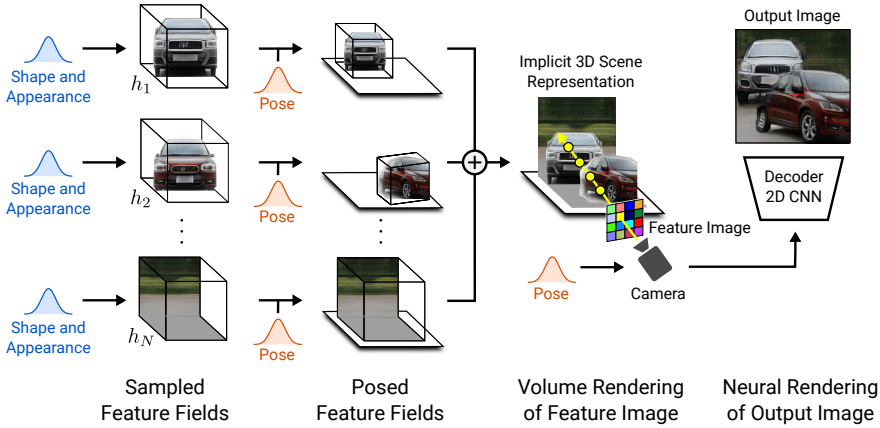
GIRAFFE

Perform volume rendering and get low-resolution feature image.



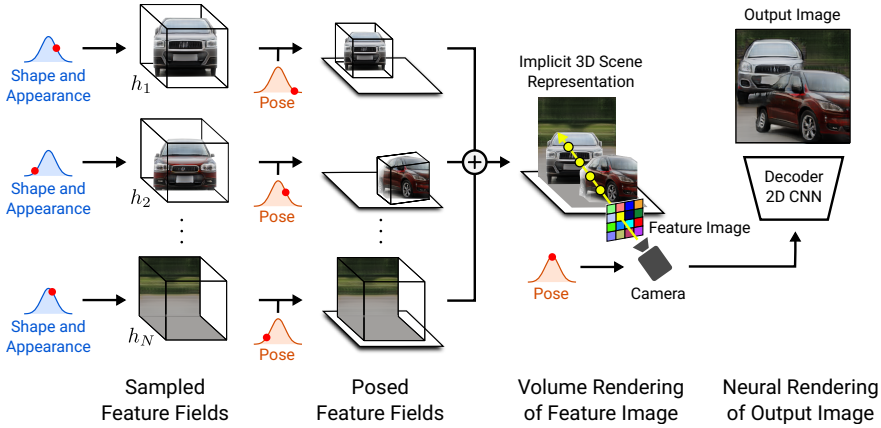
GIRAFFE

Pass feature image to neural renderer to obtain final high-resolution image.

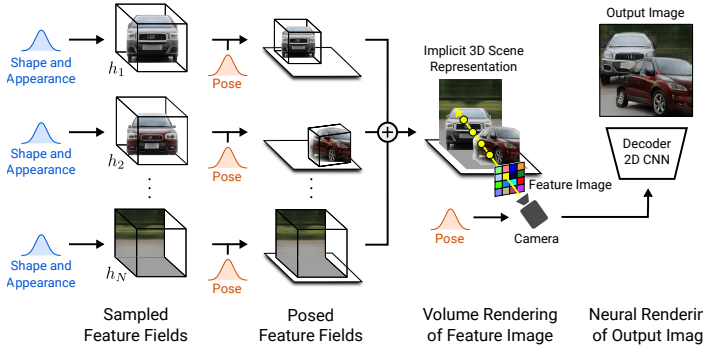


GIRAFFE

At test time, we sample individual codes to **control shapes, appearances and poses.**



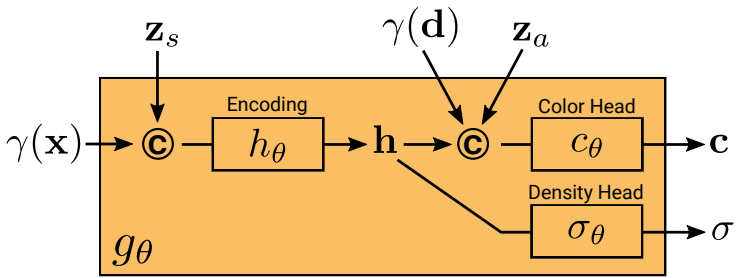
GIRAFFE



- We volume-render the feature image at 16^2 pixels and upsample via the decoder
- Unlike GRAF, we can directly train with an adversarial loss at full resolution

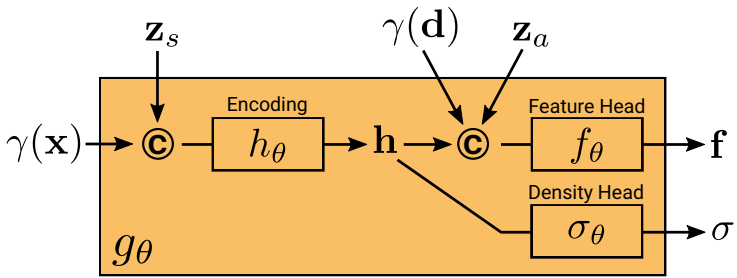
GIRAFFE

We use GRAF's architecture, but replace the color head with a **feature head:**



GIRAFFE

We use GRAF's architecture, but replace the color head with a **feature head:**



Scene Composition

We have N feature fields

$$h_i(\mathbf{x}, \mathbf{d}) = (\sigma_i, \mathbf{f}_i)$$

which predict a density σ_i and a feature vector $\mathbf{f}_i$ for any $(\mathbf{x}, \mathbf{d})$.

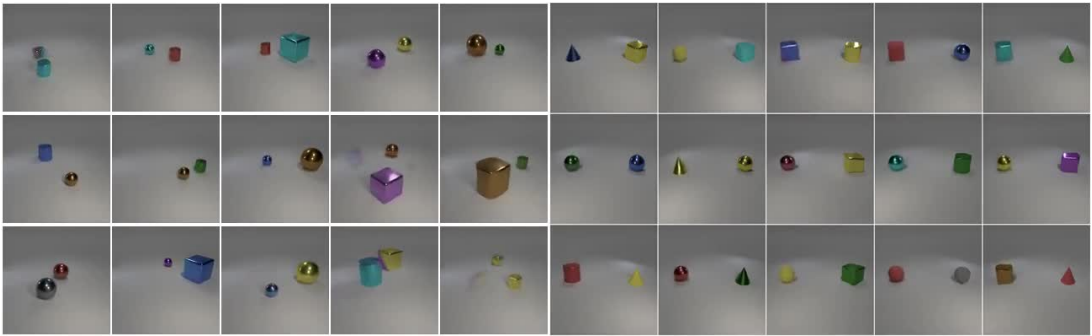
We combine densities by summation and features by density weighted averaging:

$$\sigma = \sum_{i=1}^{N} \sigma_i \qquad \mathbf{f} = \frac{1}{\sigma} \sum_{i=1}^{N} \sigma_i \mathbf{f}_i$$

Results

GIRAFFE

We compare object translation for a 2D-based GAN (left) and our method (right):



GIRAFFE

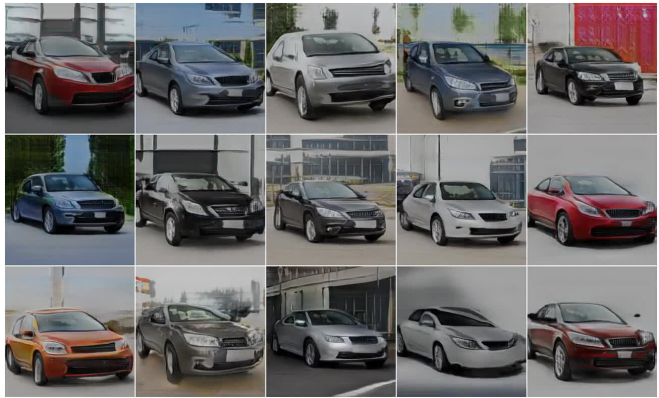
We can perform more complex operations like circular translations

GIRAFFE

We can add more objects at test time (trained on two-object)

GIRAFFE

We can rotate the object



GIRAFFE

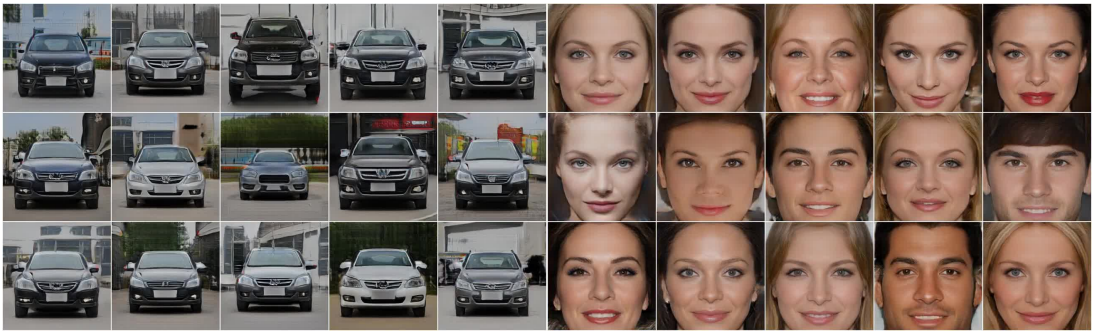
We can translate the object

GIRAFFE

We can change the object shape

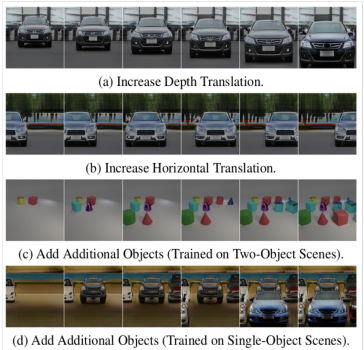
GIRAFFE

We can change the object appearance



GIRAFFE

We can generate out-of-distribution samples


(a) Increase Depth Translation.

(b) Increase Horizontal Translation.

(c) Add Additional Objects (Trained on Two-Object Scenes).

(d) Add Additional Objects (Trained on Single-Object Scenes).

Rendering Time

	64×64	256×256
GRAF	110.1ms	1595.0ms
GIRAFFE	4.8ms	5.9ms

- We volume-render the feature image at 16×16 pixels.
- CNN-based neural rendering and upsampling yields fast inference.

Can we learn the camera distribution?

CAMPARI: Camera-Aware Generative Neural Radiance Fields

GRAF, GIRAFFE:

- Learn a 3D-aware image generator from unposed image collections
- Requires **careful tuning** of **camera pose distributions**

CAMPARI: Camera-Aware Generative Neural Radiance Fields

GRAF, GIRAFFE:

- Learn a 3D-aware image generator from unposed image collections
- Requires **careful tuning** of **camera pose distributions**

CAMPARI:

- Learn a 3D aware image generator and a **camera generator** jointly.

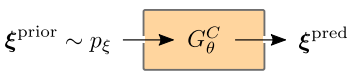
CAMPARI

Sample prior camera $\boldsymbol{\xi}^{\text{prior}} \sim p_{\xi}$. Parameters: intrinsics f_x, f_y and extrinsics r, α_r, α_e.

$$\boldsymbol{\xi}^{\text{prior}} \sim p_{\xi}$$

CAMPARI

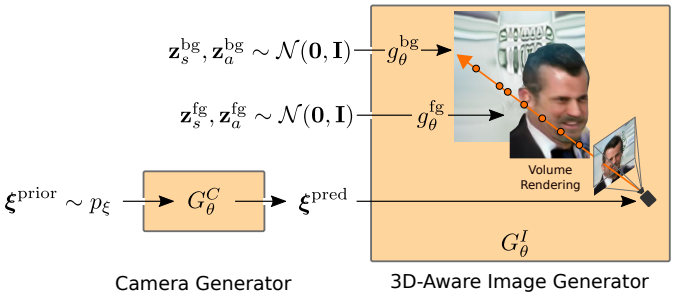
Pass $\boldsymbol{\xi}^{\text{prior}}$ to camera generator G_θ^C and obtain predicted camera $\boldsymbol{\xi}^{\text{pred}}$.



Camera Generator

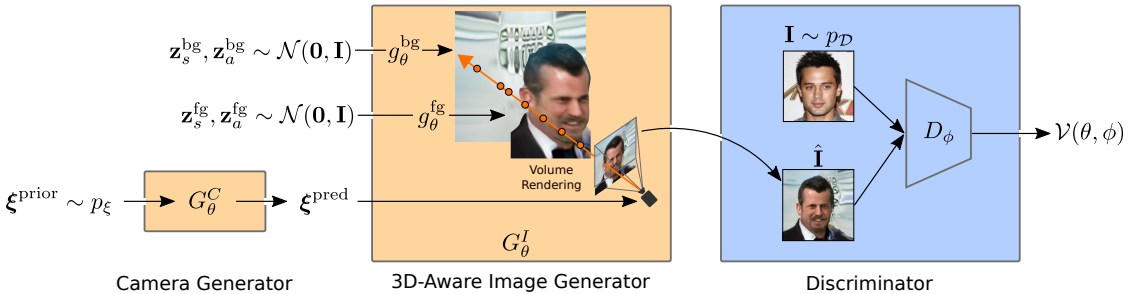
CAMPARI

Pass $\boldsymbol{\xi}^{\text{pred}}$ and sampled FG / BG latent codes to 3D-aware image generator.



CAMPARI

Train entire method with adversarial objective similar to GRAF, GIRAFFE.

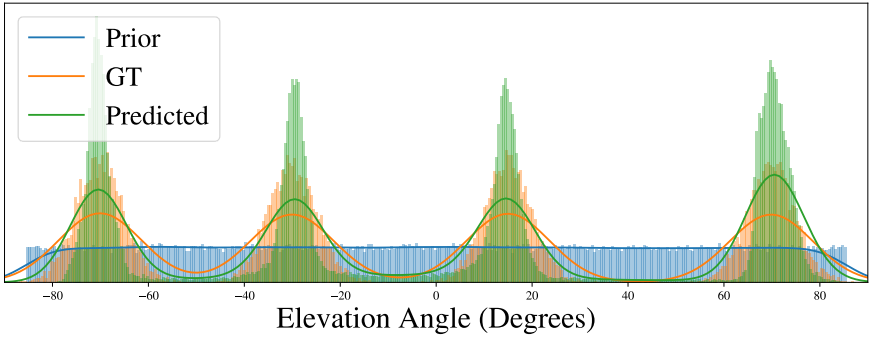


Niemeyer, Geiger: CAMPARI: Camera-Aware Decomposed Generative Neural Radiance Fields. 3DV, 2021.

Results

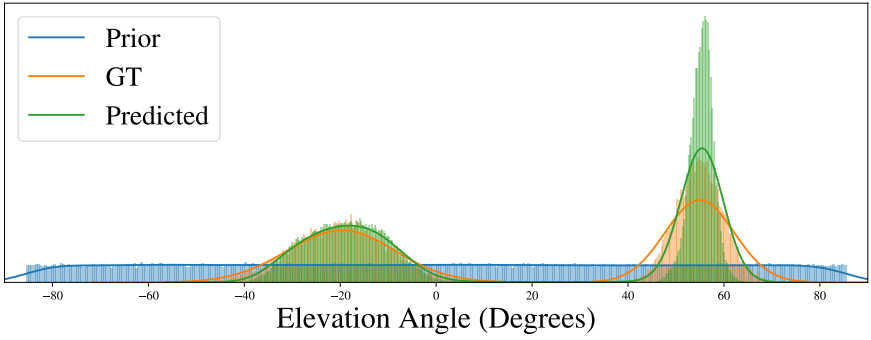
CAMPARI

CAMPARI learns to match the GT distribution for synthetic datasets.



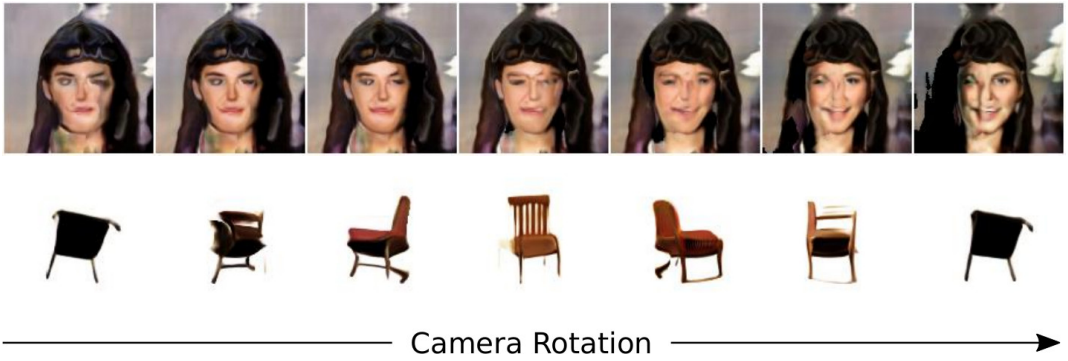
CAMPARI

CAMPARI learns to match the GT distribution for synthetic datasets.



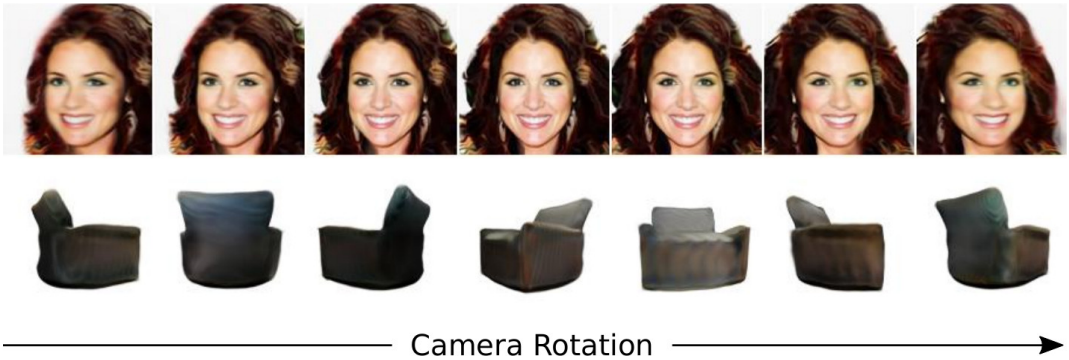
CAMPARI

This leads to significantly improved performance compared to untuned models.



CAMPARI

This leads to significantly improved performance compared to untuned models.



Generative Neural Scene Representations

Summary

- We propose novel methods for **3D controllable image synthesis**
- We demonstrate training from **raw, unposed image collections**
- We incorporate **compositional 3D scene structure** into the generative model
- We gain explicit control over **individual objects** during synthesis

Future Research

- Scale to **larger and more complex multi-object scenes**
- Enhance **photorealism** to match 2D GANs (e.g., StyleGAN2)
- Disentangle **lighting, materials,** etc.

Thank you!

http://autonomousvision.github.io

Federal Ministry
of Education
and Research

Federal Ministry
for Economic Affairs
and Energy

Microsoft
Research