GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis

Katja Schwarz. Yiyi Liao, Michael Niemeyer, Andreas Geiger

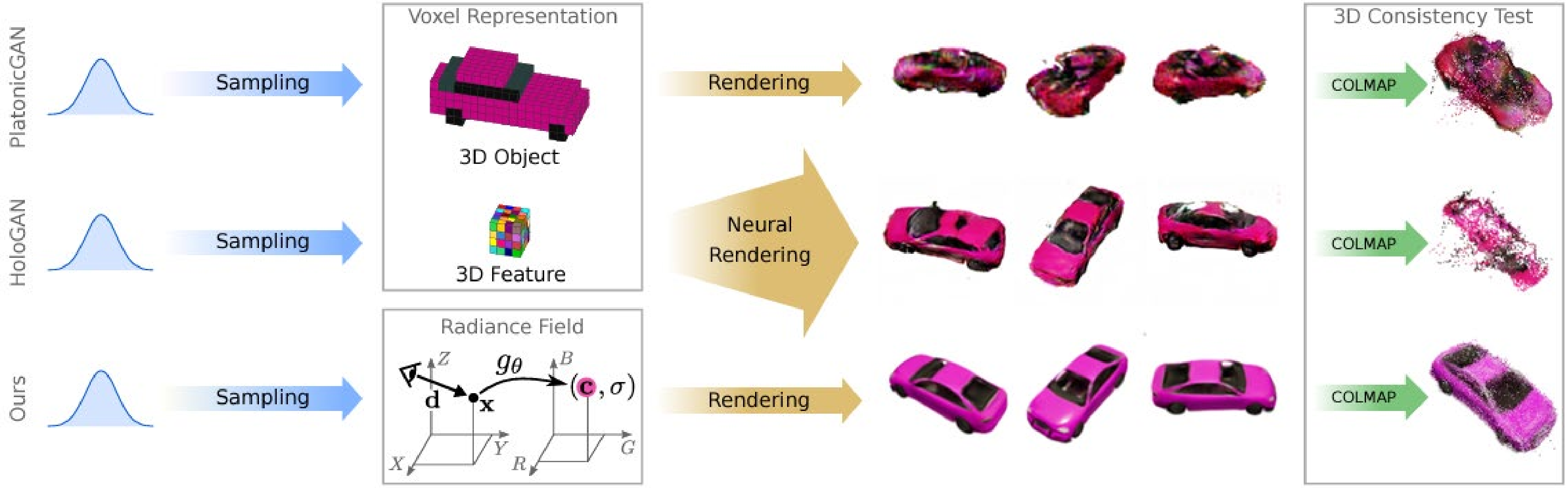
Presented by Vincent Li

Abstract

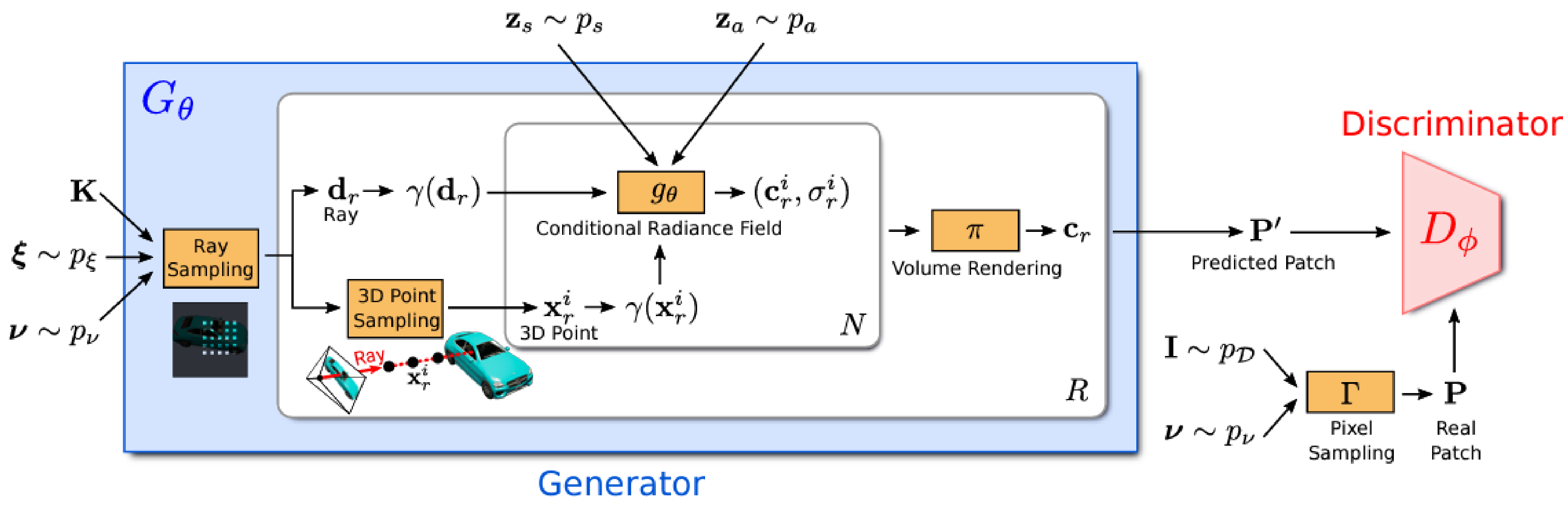
- A generative model for radiance fields for high-resolution 3D-aware image synthesis from unposed images.
  - Yield a full probabilistic generative model for drawing unconditional random samples
  - Learning from only 2D images without 3D supervision
  - Doesn’t need to be retrained for new scene (different from NeRF)
- A patch-based discriminator that samples the image at multiple scales (key to learn high-resolution generative radiance fields efficiently)
- Systematically evaluate our approach on synthetic and real datasets
  - By running a multi-view stereo algorithm (COLMAP) on several outputs to verify 3D consistency

Method Overview

- The scene is represented as a continuous function gθ that maps a location x and viewing direction d to a color value c and a volume density σ.



Generator



- camera matrix K, camera pose ξ, 2D sampling pattern v and shape/appearance codes $z_s \in R^m/z_a \in R^n$ as input and predicts an image patch P
- K is chosen in a way such that the principle point is in the center of the image

Ray Sampling

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

Figure 3: **Ray Sampling.** Given camera pose $\boldsymbol{\xi}$, we sample rays according to $\boldsymbol{\nu} = (\mathbf{u}, s)$ which determines the continuous 2D translation $\mathbf{u} \in \mathbb{R}^2$ and scale $s \in \mathbb{R}^+$ of a $K \times K$ patch. This enables us to use a convolutional discriminator independent of the image resolution.

Conditional Radiance Field

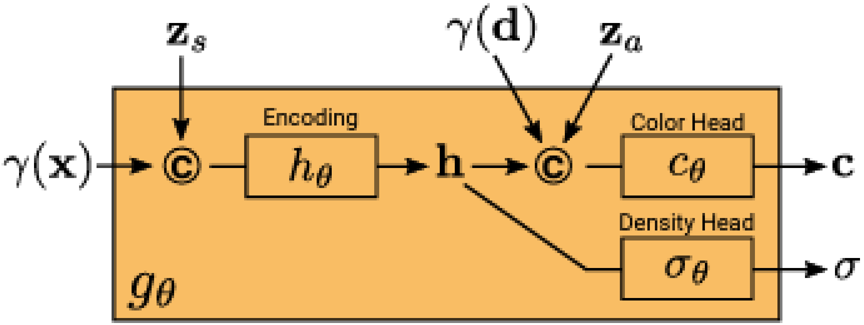


Figure 4: **Conditional Radiance Field.** While the volume density σ depends solely on the 3D point $\mathbf{x}$ and the shape code $\mathbf{z}_s$, the predicted color value $\mathbf{c}$ additionally depends on the viewing direction $\mathbf{d}$ and the appearance code $\mathbf{z}_a$, modeling view-dependent appearance, e.g., specularities.

- Where the network is :)
- In contrast to NeRF, CRF is also conditioned on shape code z_s and z_a in addition to position x and viewing direction d
- σ is computed independently of the view point d and appearance code to disentangle shape and appearance.

Discriminator

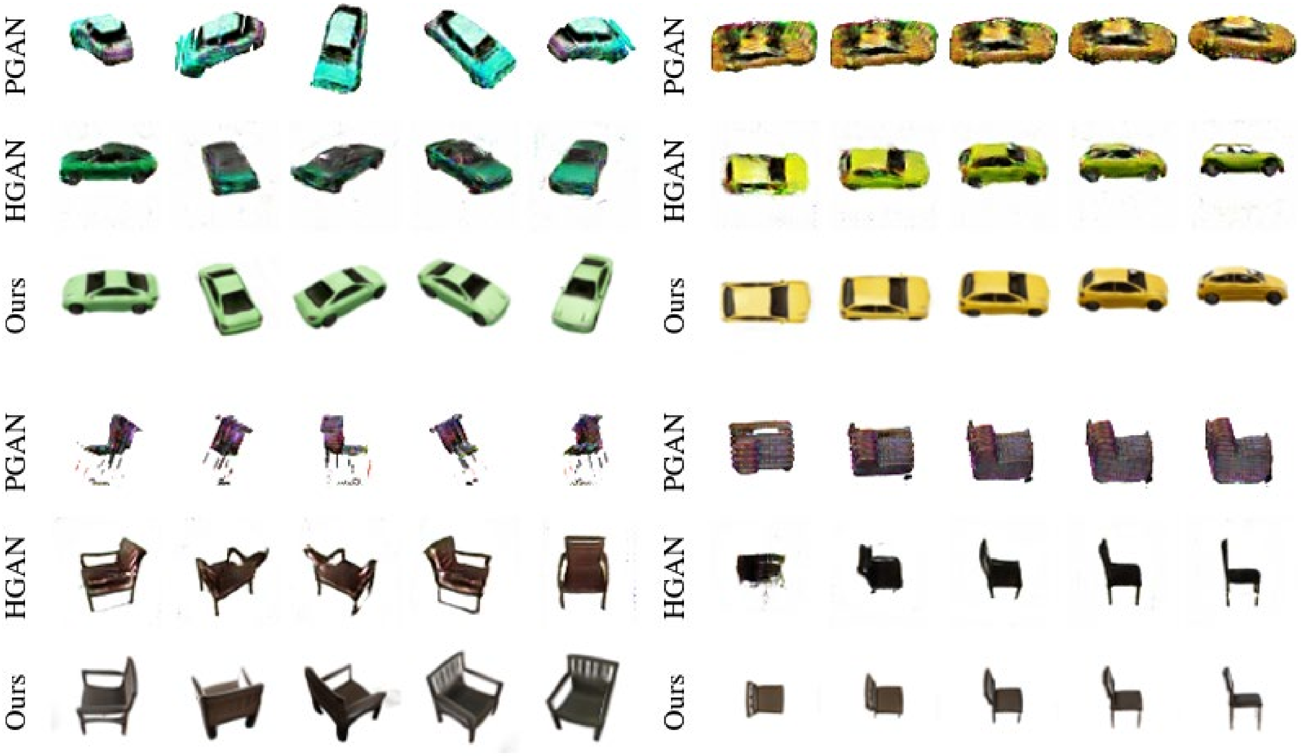
- A K x K patch is extracted from the real image using a v ~ p_v (same as generator)
- Then sample the real patch P by querying I at the 2D image coordinates P(**u**, s) using bilinear interpolation.
- Very similar to PatchGAN however continuous displacement **u** and scale sis allowed while PatchGAN uses s = 1.
- Noted that real image **I** is not downsampled, but queried using sparse locations to retain high-frequency details

Training and Inference

$$
\begin{aligned}
V(\theta, \phi) = & \ \mathbb{E}_{\mathbf{z}_s \sim p_s, \mathbf{z}_a \sim p_a, \boldsymbol{\xi} \sim p_\xi, \boldsymbol{\nu} \sim p_\nu} \left[f(D_\phi(G_\theta(\mathbf{z}_s, \mathbf{z}_a, \boldsymbol{\xi}, \boldsymbol{\nu}))) \right] \\
& + \mathbb{E}_{\mathbf{I} \sim p_\mathcal{D}, \boldsymbol{\nu} \sim p_\nu} \left[f(-D_\phi(\Gamma(\mathbf{I}, \boldsymbol{\nu}))) - \lambda \|\nabla D_\phi(\Gamma(\mathbf{I}, \boldsymbol{\nu}))\|^2 \right]
\end{aligned}
$$

- Non-saturating GAN with R1-regularization

Results



(a) Rotation (b) Elevation

Results

- How do Generative Radiance Fields compare to voxel-based approaches?

	Chairs	Birds	Cars	Cats	Faces
2D GAN [35]	59	24	66	18	15
PLATONICGAN [20]	199	179	169	318	321
HoloGAN [40]	59	78	134	27	25
Ours	34	47	30	26	25

Table 1: **FID** at image resolution 64^2 pixels.

Results

- Do 3D-aware generative methods scale to high - resolution outputs?

	Cars			Faces		
	128	256	512	128	256	512
HoloGAN [40]	211	230	–	39	61	–
w/o 3D Conv	180	189	251	31	33	51
Ours	41	71	84	35	49	49
upsampled	–	91	128	–	63	77
sampled	–	74	104	–	50	56

Table 2: **FID** at image resolution 128^2-512^2.

Results

- Should learned projections be avoided?

Figure 6: **Viewpoint Interpolations** on Faces and Cars at image resolution 256^2 pixels for HoloGAN [40] (HGAN), HoloGAN w/o 3D Conv (HGAN ~~3D~~) and our approach (Ours).

Results

- Continued

Figure 7: **3D Reconstruction** from synthesized images at resolution 256^2. Each pair shows one of the generated images and the 3D reconstruction from COLMAP [61].

Method	MMD-CD
Ours	**0.044**
HGAN	0.109
HGAN ~~3D~~	0.092

Table 3: **Reconstruction Accuracy** on Cars for 100 COLMAP reconstructions compared to their closest shapes in the ground truth in terms of MMD [1] measuring chamfer distance (CD).

Results

- Are Generative Radiance Fields able to disentangle shape from appearance?

Figure 8: **Disentangling Shape / Appearance.** Results from our model on Cars, Chairs and Faces.

Limitation

- Simple scenes with single objects