Generative Occupancy Network for 3D Surface-Aware Image Synthesis

Anonymous Author(s)
Affiliation
Address
email

Abstract

The advent of generative radiance field has significantly promoted the development of 3D-aware image synthesis. The cumulative rendering in radiance field scatters the gradients along the ray and empowers the application in hard-training generative models, but prompts the diffused surface of generated objects. At the same time, occupancy network predicts a deterministic surface inherently. Nevertheless, it always requires precise object masks for training and, more importantly, discourage the convergence of generative models especially in the early stage. In this paper, we propose **Generative Occupancy Network** (GON), which leverages a dedicated transition, *i.e.*, *start from radiance field, march to occupancy representation*. Specifically, the initially sampled points approximately cover the entire volume, thus facilitating the optimization and providing a proper warm-up for the shape. As the training goes, these points approach the surface gradually until a minimal neighbor region, ensuring an accurate surface eventually. Exhaustive experiments on multiple datasets demonstrate that GON can synthesize high-fidelity images with 3D consistency and recover the compact surface simultaneously.

1 Introduction

Deep generative adversarial networks [1, 2, 3, 4] have demonstrated their superiority of synthesizing photorealistic and striking images. However, these models are constraint in 2D domain, struggling to generate 3D consistent images, let alone recovering the underlying 3D object shape. Despite similarly learning from unposed image collections only, 3D-aware image synthesis is more appealing since it intends to learn a 3D scene representation directly. Hence, multi-view 3D consistent images can be generated by manually controlling the rendered camera poses. Moreover, an explicit 3D shape can be recovered from, which paves the way for various downstream tasks such as shape editing and relighting.

Owing to its flexibility, neural radiance field enables impressive novel view syntheis [5], and is further applied to 3D-aware generative models, namely generative radiance field (GRAF). Note that, the volume rendering [6] used in radiance field holds the superior ability for rendering translucent objects intrinsically as it accumulates the colored desities along the ray. Accordingly, the successful employment in hard-training generative models is inextricable from dispersed gradients distributed in the whole volume. Such a characteristic permits the optimization of generative models and gives rise to 3D consistent image synthesis, but still incurs an inevitable incapacity for ensuring an accurate object surface. Definitely, the recovered diffused surface significantly hampers the succeeding applications.

In contrast to radiance field, occupancy network [9] manages to yield a compact surface due to its intentional design. As per the definition in [9], the 3D surface is implicitly represented by

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

the continuous decision boundary of a deep neural network classifier, guaranteeing the surface compactness naturally.

2 Related Work

3 Methodology

3.1 Diffused Surface of Generative Radiance Fields

Neural radiance fields (NeRF) [5] has been applied to novel view synthesis and achieves remarkable performance on complex scenes. NeRF represents a static scene as volume densities and view-dependent RGB colors. Specifically, given a 3D point $\mathbf{x} \in \mathbb{R}^3$ in space and a view direction $\mathbf{d} \in \mathbb{R}^3$, NeRF model will predict the volume density $\sigma(\mathbf{x}) \in \mathbb{R}$ and the emitted color $c(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^3$. For the rendering, NeRF leverages the classic volume rendering technique to accumulate the colored densities of N sampled points $\{\mathbf{x}_i = \mathbf{o} + t_i\mathbf{d}\}$ along a camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, where $\mathbf{o}$ stands for the camera origin. Eventually, the integrated color is estimated via alpha composition as follows:

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{i=1}^{N} T_i \Big(1 - \exp\big(-\sigma(\mathbf{x}_i)\delta_i\big)\Big) c(\mathbf{x}_i, \mathbf{d}), \text{ where } T_i = \exp(-\sum_{j=1}^{i-1} \sigma(\mathbf{x}_j)\delta_j), \tag{1}$$

where $\delta_i = |x_{i+1} - x_i|$ is the distance between adjacent points. Note that, equation (1) is naturally differentiable and the model could be directly optimized via back-propagation.

By training on posed 2D images, NeRF can synthesize photorealistic and consistent novel views. Instead, generative radiance fields (GRAF) leverages the unposed 2D images corpus for 3D-aware image synthesis. Recently, several works [7, 8] have made great stride in generative radiance fields and can generate high-fidelity images with 3D consistency. The cumulative rendering in GRAF considers multiple points on a ray, making these generative models easy to optimize, but failing to maintain the surface compactness of generated objects. As illustrated in Fig1(a)[R], the state-of-the-art model π-GAN is prone to predict diffused surface, where the volume density smoothly spread around the surface position. As a consequence, GRAF has to sample dozens of points rather than one for each ray during inference, which brings huge inconvenience for real-time rendering. Besides, the lack of an explicit surface during training making it hard to infer the precise 3D object shape, as shown in Fig1(a).

3.2 From Generative Radiance Fields to Generative Occupancy Network

Unlike radiance field possessing a diffused surface, occupancy network [9] intentionally models an explicit surface. For a point $\mathbf{x} \in \mathbb{R}^3$ in 3D space, the corresponding occupancy o_θ and color c_θ are both described with neural networks:

$$\begin{aligned} o_\theta(\mathbf{x}) &: \mathbb{R}^3 \to [0, 1] \\ c_\theta(\mathbf{x}) &: \mathbb{R}^3 \to \mathbb{R}^3. \end{aligned} \tag{2}$$

According to the definition, the 3D surface of an object is determined by the level set $o_\theta(\mathbf{x}) = \tau$ for a predefined threshold $\tau \in [0, 1]$. After training, occupancy network is able to learn continuous occupancy, insuring to recover a compact decision boundary, *i.e.*, the object surface. A plethora of works [10, 11, 12, 13] have introduced occupancy representation to 3D object reconstruction by making the rendering process differentiable and shown impressive performance. However, occupancy network always entails precise object masks to capture accurate geometry [12, 14]. More importantly, only the surface point is considered for rendering, which results in sparse gradients for each ray and impedes the convergence of applied generative models. Considering that the radiance field representation encourages the training, a nontrivial combination of both, *i.e.* "start from radiance fields, march to occupancy representation" can have your cake and eat it too.

In this paper, we propose **Generative Occupancy Network** (GON), which capitalizes on the transition above along the whole training procedure. While a concurrent work [14] adopts similar idea for 3D reconstruction of static scene, our approach targets learning generative models from unposed images, which is more challenging. (xxd: put in the related work?) Specifically, GON always finds the surface point $\mathbf{x}_s$ firstly via root-finding [10] and sample points only in the surface neighbor

regions. For a ray $r(t) = \mathbf{o} + t\mathbf{d}$, let $\mathbf{x}_s = \mathbf{o} + t_s\mathbf{d}$ denote the surface point, and N points $\{\mathbf{x}_i\}$ are obtained with stratified sampling within the neighbor region $[t_s - \Delta, t_s + \Delta]$:

$$t_i \sim \mathcal{U}\left[t_s - \Delta + \frac{2i}{N}\Delta, t_s - \Delta + \frac{2i+2}{N}\Delta\right], \text{ where } i = 0, 1, ..., N-1. \tag{3}$$

As illustrated in Fig2 [R], the initial neighbor interval Δ is around a half of the distance between near t_n and far bounds t_f. Hence, the sampled points at this stage almost cover the entire volume, leading to an appropriate initialization of coarse shapes. As the training goes, the predicted shape turns more and more accurate hence we shrink the neighbor regions gradually. In detail, the sampling interval Δ will decrease monotonically with an exponential decay rate γ until it drops to a pre-defined minimal value Δ_{min}, where all the points are near the surface. Without bells-and-whistles, the proposed transition ensures the surface smoothness of generated 3D objects as shown in Fig1(b) [R].

In order to unify the format of radiance field and occupancy, volume rendering equation (1) can be reformulated to:

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{i=1}^{N} o(\mathbf{x}_i) \prod_{j<i} \big(1 - o(\mathbf{x}_j)\big) c(\mathbf{x}_i, \mathbf{d}), \tag{4}$$

where $o(\mathbf{x}_i) = 1 - \exp(-\sigma(\mathbf{x}_i)\delta_i) \in [0, 1]$ can be interpreted as the occupancy. In the experiments, we predict $o(\mathbf{x})$ directly instead of density $\sigma(\mathbf{x})$.

3.3 Training and Inference

Following π-GAN [7], we randomly sample camera poses ξ from a dataset-related distribution p_ξ. Denote I as a real image from data distribution $p_{\mathcal{D}}$ and the non-saturating GAN loss can be described as follows:

$$\begin{aligned} \mathcal{L}_{\text{origin}}(\theta, \phi) = \mathbf{E}_{\mathbf{z}\sim p_{\mathbf{z}}, \xi\sim p_\xi} &\left[f\Big(D_{\theta_D}(G_{\theta_G}(\mathbf{z}, \xi))\Big)\right] \\ &+ \mathbf{E}_{I\sim p_{\mathcal{D}}} \left[f(-D_{\theta_D}(I)) + \lambda |\nabla D_{\theta_D}(I)|^2 \right], \\ &\text{where } f(u) = -\log(1 + \exp(-u)). \end{aligned} \tag{5}$$

However, only with $\mathcal{L}_{\text{origin}}$ to guide the training, generated shapes tend to be messy and contain smoke-like artifacts. To reduce such artifacts and make the surface more smooth, we extra incorporate two priors.

The first introduced prior is surface normal regularization, which enforces more natural shapes [10]:

$$\mathcal{L}_{\text{normal}} = \sum_{\mathbf{x}_s \in \mathcal{S}} ||\mathbf{n}(\mathbf{x}_s) - \mathbf{n}(\mathbf{x}_s + \boldsymbol{\epsilon})||_2. \tag{6}$$

Here, $\boldsymbol{\epsilon}$ is a small random 3D perturbation and $\mathbf{n}$ denotes the normal vector, which can be computed by $\mathbf{n}(\mathbf{x}) = \nabla_{\mathbf{x}} o_\theta(\mathbf{x}) / ||\nabla_{\mathbf{x}} o_\theta(\mathbf{x})||_2$.

The second term is opacity prior aiming to reduce the entropy of the predicted occupancy $o(\mathbf{x})$:

$$\mathcal{L}_{\text{opacity}} = \frac{1}{N} \sum_{i=1}^{N} \log(o(\mathbf{x}_i)) + \log(1 - o(\mathbf{x}_i)). \tag{7}$$

Therefore, the final loss for training can be written as:

$$\mathcal{L}(\theta, \phi) = \mathcal{L}_{\text{origin}}(\theta, \phi) + \lambda_{\text{normal}} \mathcal{L}_{\text{normal}} + \lambda_{\text{opacity}} \mathcal{L}_{\text{opacity}}, \tag{8}$$

where λ_{normal} and λ_{opacity} are both loss weights.

At inference time, the generator will sample points in the minimal interval Δ_{min} and integrate the final color with equation (4). Note that, we can only use the surface point on each ray to do the rendering. As shown in Fig1(b)[R], the image rendered with only one point is almost indistinguishable from that with multiple points. Such equivalence can be guaranteed theoretically when $\Delta_{\text{min}} \to 0$. Please refer to the supplementary material for the formal proof.

3.4 Implementation Details

4 Experiments

5 Conclusion

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[2] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," in *International Conference on Learning Representations*, 2018.

[3] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

[4] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

[5] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, pp. 405–421, Springer, 2020.

[6] J. T. Kajiya and B. P. Von Herzen, "Ray tracing volume densities," *ACM SIGGRAPH computer graphics*, vol. 18, no. 3, pp. 165–174, 1984.

[7] E. R. Chan, M. Monteiro, P. Kellnhofer, J. Wu, and G. Wetzstein, "pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis," in *CVPR*, 2021.

[8] K. Schwarz, Y. Liao, M. Niemeyer, and A. Geiger, "Graf: Generative radiance fields for 3d-aware image synthesis," in *NeurIPS*, vol. 33, 2020.

[9] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4460–4470, 2019.

[10] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger, "Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3504–3515, 2020.

[11] Y. Jiang, D. Ji, Z. Han, and M. Zwicker, "Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[12] L. Yariv, Y. Kasten, D. Moran, M. Galun, M. Atzmon, B. Ronen, and Y. Lipman, "Multiview neural surface reconstruction by disentangling geometry and appearance," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[13] S. Liu, Y. Zhang, S. Peng, B. Shi, M. Pollefeys, and Z. Cui, "Dist: Rendering deep implicit signed distance function with differentiable sphere tracing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2019–2028, 2020.

[14] M. Oechsle, S. Peng, and A. Geiger, "Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction," *arXiv preprint arXiv:2104.10078*, 2021.