Zero-Shot Text-Guided Object Generation with **Dream Fields**

Ajay Jain Ben Mildenhall Jon Barron Pieter Abbeel Ben Poole
ajayj@berkeley.edu, {bmild, barron}@google.com, pabbeel@berkeley.edu, pooleb@google.com

Google Research

text-to-3D in Colab!
ajayj.com/dreamfields

Berkeley UNIVERSITY OF CALIFORNIA

Text-to-image generation now supports incredible quality and diversity.

"an armchair in the shape of an avocado." (DALL-E)

"A chromeplated cat sculpture placed on a Persian rug." (Imagen)

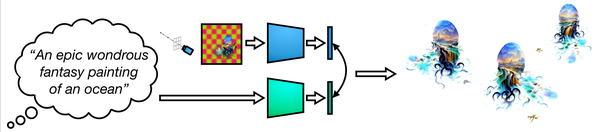
But due to limited data, **generative models for 3D are narrow and category specific**.

Neural Radiance Fields are photorealistic, but need many multi-view photo captures. How to design novel content?

Captured images + poses → **Photometric loss** → **Novel view**

$$\mathbf{C}(r,\theta) = \int_{t_n}^{t_f} T(\mathbf{r},t)\sigma_\theta(\mathbf{r}(t))\mathbf{c}_\theta(\mathbf{r}(t))dt$$

Can 3D reconstruction systems be trained with **no photos**? We perform **text-to-3D synthesis** by training Neural Radiance Fields with ***zero input photos, only a caption.***



Problems / Contributions

1) Lack of 3D data

Problem: There aren't many 3D models available, making it difficult to train 3D generation.

Solution: **Captioned photographs are abundant** and diverse. We repurpose **scalable pre-trained image-text representations like CLIP**.

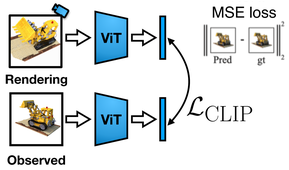
2) Underconstrained scene representation

Problem: Photorealistic 3D scene representations like NeRF are too flexible without photos or learned priors.

Solution: Regularize scene to be **opaque but compact**. Improve MLP for easier optimization.

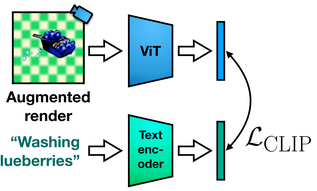
Training Neural Radiance Fields with CLIP

DietNeRF regularized NeRF with a feature space loss, supervising novel views during training and **transferring 2D knowledge.**



Intuition: **All views should have consistent semantics.** "A bulldozer is a bulldozer from any perspective."

Our insight: **Train NeRF entirely in feature space.** An **aligned image-text representation** like CLIP allows us to **control synthesis with a caption.**



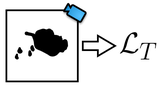
$$\mathcal{L}_{\text{CLIP}}(\theta, \text{pose } \mathbf{p}, \text{caption } y) = -\mathbf{g}(\mathbf{I}(\theta,\mathbf{p}))^{\mathrm{T}}\mathbf{h}(\mathbf{y})$$

Regularizing Dream Fields

Dream Fields lacks a 3D prior, and NeRF can fit highly degenerate geometry (floaters, occlusions, translucency).

To fix geometry, we remove view dependence to model the shape without lighting and **regularize mip-NeRF to be sparse and compact**.

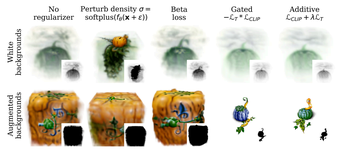
Augment backgrounds + random crop



+ Maximize transmittance

$$T(\mathbf{r},\theta,t) = \exp\left(-\int_{t_n}^{t}\sigma_\theta(\mathbf{r}(s))ds\right)$$

$$\mathcal{L}_T = -\min(\tau, \text{mean}(T(\theta,\mathbf{p})))$$



"an illustration of a pumpkin on the vine."

Text to 3D with multi-view consistency

bouquet of flowers sitting in a clear glass vase.

a small green vase displays some small yellow blooms.

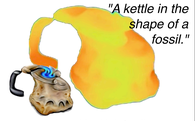
a slug crawling on the ground around flower petals.

Compositional 3D generation --> easy concept exploration.

"an armchair/teapot/snail in the shape of a _____."

Prompt template: an armchair in the shape of a ____, an armchair imitating a ____.

Object centric prompts		R-Precision ↑ CLIP B/32	R-Precision ↑ LiT_{uu} B/32
Baseline	COCO GT images	77.1±3.4	75.2±3.5
	Simplified NeRF	31.4±2.7	10.8±1.8
Positional encoding	+ mip-NeRF IPE	29.7±2.6	12.4±1.9
	+ Higher freq. Fourier features	24.2±2.5	10.5±1.8
Sparsity, augment	+ random crops	25.8±2.5	10.5±1.8
	+ transmittance loss	23.7±2.4	7.6±1.5
	+ background aug.	44.1±2.8	26.1±2.5
Scene param.	+ MLP architecture	52.0±2.9	27.8±2.6
	+ scene bounds	65.4±2.7	**38.9±2.8**
	+ track origin	59.8±2.8	34.6±2.7
Scaling	+ LiT_{uu} ViT B/32	59.5±2.8	–
	+ 20K iterations, 252^2 renders	**68.3±2.7**	–

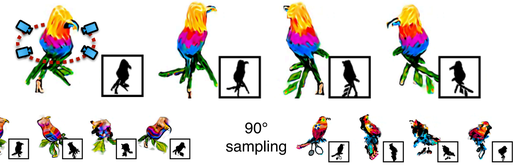


(b) Neural Radiance Fields

(c) Dream Fields

Camera sampling

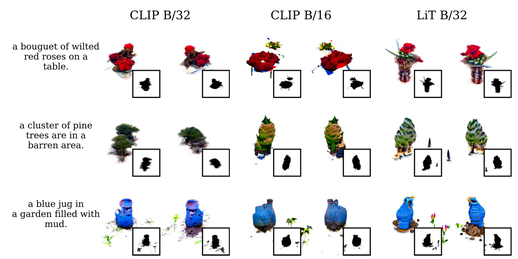
"A bird that has many colors on it."



Diverse generations: changing the seed

Representation matters: Scaling image-text data

Optimized model	Retrieval model R-Precision CLIP B/32	CLIP B/16	LiT_{uu} B/32
COCO GT	77.1±3.4	79.1±3.3	75.2±3.5
CLIP B/32 [46]	*(86.6±2.0)*	74.2±2.5	42.8±2.8
CLIP B/16 [46]	59.8±2.8	*(93.5±1.4)*	35.6±2.7
LiT_{uu} B/32	59.5±2.8	66.7±2.7	*(88.9±1.8)*



Limitations and future work

"A small surfboard sign that says TRADER VIC'S, Los Angeles."

- Only optimize individual viewpoints.
- Limited by CLIP / LiT training data
- Quality varies across prompts. Complex prompts are challenging.
- Improving realism, text.
- Exporting assets for graphics.

NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, Mildenhall, Srinivasan, Tancik, et al, ECCV 2020
Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields, Barron et al, ICCV 2021
CLIP: Learning Transferable Visual Models From Natural Language Supervision, Radford et al 2021.
LiT: Zero-Shot Transfer with Locked-image text Tuning, Zhai et al CVPR 2022
Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis, Jain et al, ICCV 2021
DALL-E: Zero-Shot Text-to-Image Generation, Ramesh et al, 2021
Differentiable Image Parameterizations, Mordvintsev, et al., Distill, 2018.