RenderFormer V2

Neural Rendering with Heterogeneous Scene Primitives

A pretrained transformer for turning a sequence of mixed scene primitives into a globally illuminated image.

Chong Zeng[1] **Yue Dong**[2] **Pieter Peers**[3] **Lvmin Zhang**[1] **Maneesh Agrawala**[1]

[1]Stanford University [2]Microsoft Research [3]College of William & Mary

ECCV 2026 | MALMÖ | SEPT 8–12

PAPER · CODE · MODEL · GALLERY

HIGHLIGHTS

128K
primitives in one forward pass. RenderFormer-V1 stops at 4K

2048²
output pixels, without upsampling stage

+9.9 dB
PSNR increase over RenderFormer-V1 at 128K triangles, with 4.8× lower LPIPS

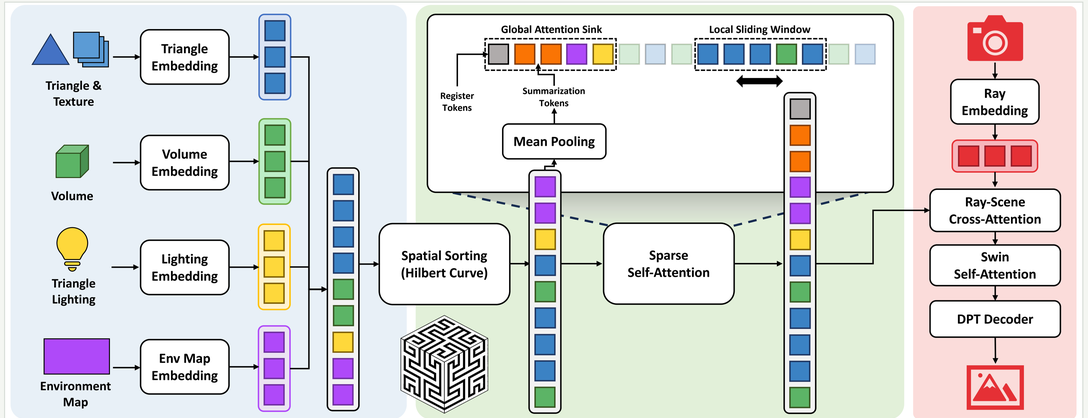
3.5×
faster than RenderFormer-V1 and 4.7× faster than Cycles

No
per-scene training, fine-tuning or hand-written transport code

01 / ARCHITECTURE

A scene is a token sequence.

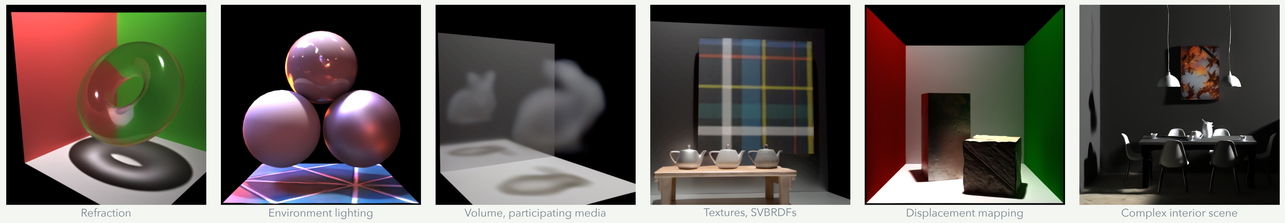
- Geometry, participating media, area lights, environment maps and camera rays are each represented as tokens of one sequence
- View-independent stage resolves primitive-to-primitive transport; view-dependent stage turns ray bundles into pixel values



02 / COVERAGE

One model, many light transport effects.

- Refraction, environment lighting, volumetric scattering, textures and displacement, modeled with a single unified architecture and weights
- No effect-specific code path: each effect is learned by example



WHAT EACH TOKEN CARRIES

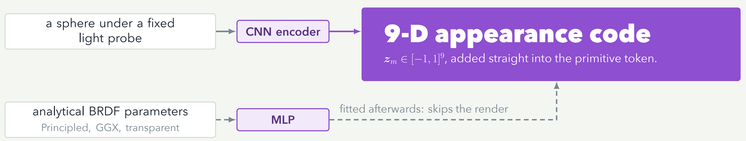
Triangle	3 vertices + 3 normals; a 9-D material code; a 32×32, 13-channel patch (9 material, 3 normal, 1 displacement) through a frozen VAE
Voxel medium	rotation + per-axis scale; a 4^3 grid of RGB scattering, RGB absorption and phase anisotropy
Area light	3 vertices + 3 normals; RGB emittance, in its own token rather than fused into geometry
Environment map	LDR and log-HDR 512×256, each through a frozen VAE into an 8×4 token sequence; direction map; log-max intensity
Camera bundle	8×8 ray directions, in world space

A single positional encoding covers all of them: RoPE on the primitive centroid, with position-less tokens falling back to the scene centroid, so the whole embedding is translation invariant.

03 / MATERIAL

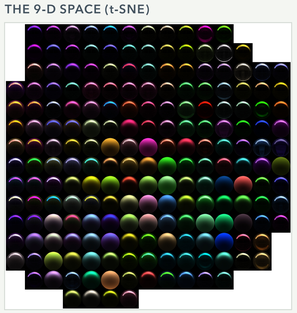
Data-driven reflectance model.

- No hard-coded reflectance model
- A 9-D latent space encodes *appearance*: a sphere under a fixed light probe, encoded by a CNN
- Never explicitly decoded; the model learns to evaluate reflectance from the latent code

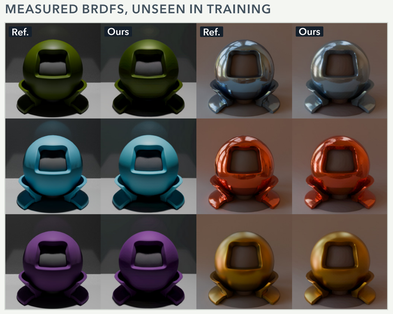


Because the space is defined by rendered appearance, any material you can render is usable, including measured BRDFs never seen in training.

THE 9-D SPACE (t-SNE)



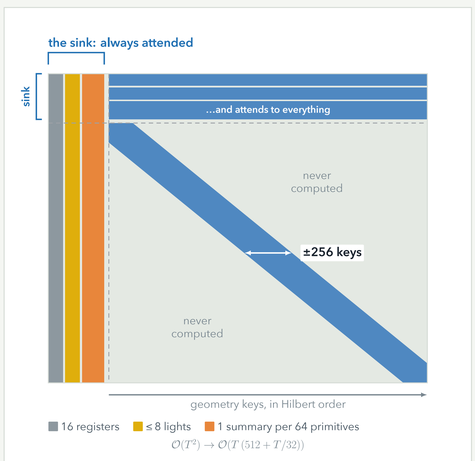
MEASURED BRDFS, UNSEEN IN TRAINING



04 / ATTENTION

What is the attention sink?

- Nearby primitives dominate transport, so each attends inside a ±256 window of the Hilbert order
- What is *not* local, such as illumination and coarse occlusion, is also included in the sink
- The sink is always attended, and attends back to every token



Every row is one query, every column one key; sink and window widths are exaggerated for legibility.

WHY HILBERT ORDER

Sorting the primitives along a Hilbert curve makes neighbours in space localized in the sequence, so a 1-D sliding window buys a spatial one.

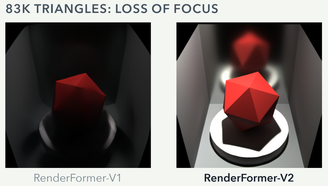
THREE KINDS OF SINK TOKENS

16 registers
Free scratch space for whatever is globally true of the scene.

Every light
Each primitive sees each light whatever the distance, which is the attention analogue of importance sampling the lights.

Summaries
Mean-pooling every 64 geometry tokens keeps long-range occlusion in reach, coarsely: distant detail does not matter.

83K TRIANGLES: LOSS OF FOCUS



RenderFormer-V1 **RenderFormer-V2**

Dense attention over 83K triangles loses focus and darkens the scene. The sink keeps the highlight and the contact shadow.

05 / EVALUATION

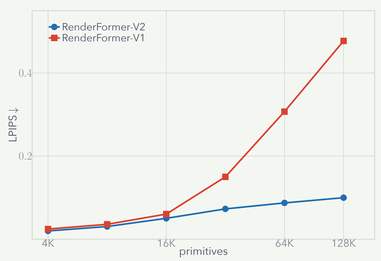
What does each part contribute?

- Held-out scenes; every variant trained through the 16K-primitive stage
- Sparse attention beats full attention, because dense attention tends to lose focus

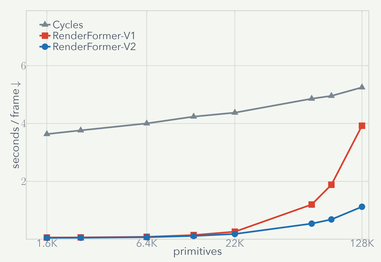
ABLATION OF THE SPARSE ATTENTION

Attention variant	PSNR↑	SSIM↑	LPIPS↓	HDR-FLIP↓
RenderFormer-V2	**28.25**	**.8982**	**.0997**	**.4200**
w/o attention sink	27.36	.8841	.1247	.4359
w/o sliding window	26.92	.8715	.1289	.4517
sink w/o lights	27.40	.8813	.1081	.4250
sink w/o summaries	27.09	.8754	.1210	.4485
sink w/o both	26.13	.8474	.1514	.4959
summaries 1 : 32	27.78	.8923	.1072	.4231
summaries 1 : 128	26.48	.8690	.1346	.4655
summaries 1 : 256	26.93	.8663	.1316	.4580
full attention	27.91	.8953	.1027	.4287

QUALITY vs SCENE SIZE



RUNTIME vs SCENE SIZE



Paper, code, pretrained model and full gallery
renderformer.github.io/v2