# RADIANT: OBJECT-CENTRIC FEED-FORWARD GLOBAL ILLUMINATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Feed-forward neural renderers predict physically grounded global illumination across unseen 3D scenes in a single forward pass, bypassing expensive Monte Carlo path tracing and per-scene optimization. However, existing methods rely on tokenizing explicit triangle meshes, causing attention complexity to scale quadratically with primitive count. Macroscopic light transport operates predominantly between surface patches and objects rather than diffusing facet by facet across individual triangles; polygon tessellation is an artifact of representation rather than transport physics. We introduce RADIANT (**RAD**iance-based **I**nter-**A**ctor **N**eural **T**ransformer), an object-centric architecture that learns mesh-free feed-forward global illumination directly from point-sampled surfaces. By pooling surface points into a small number of patch tokens per object, RADIANT decouples attention complexity from polygon discretization. Light transport is modeled in two stages: an inter-object scene transformer and a view decoder. A factored radiometric attention bias injects directional alignment and material compatibility via rotary phases without explicit pairwise matrices, while an invariant form-factor extension incorporates pairwise geometric exchange invariants directly as $\mathrm{SE}(3)$-invariant inner products. To resolve primary visibility without mesh connectivity, camera rays cross-attend scene tokens guided by depth-tested point-splat hit channels in a shared spherical frame. Across 1.03M scenes under matched capacity, RADIANT achieves 25 dB in $5\times$ fewer steps and surpasses RenderFormer by 3.0 dB (and RenderFormer-V2 by 7.2 dB) at 300k steps. On scenes with hundreds of objects, RADIANT avoids quadratic primitive bottlenecks and maintains interactive rendering rates. We will release our code and data upon publication.

## 1 INTRODUCTION

Physically based global illumination is fundamental to graphics, vision, and embodied AI. Governed by the rendering equation (Kajiya, 1986), simulating realistic light transport (soft shadows, diffuse interreflections, and color bleeding) is essential for photorealism and spatial perception. While classical path tracing (Kajiya, 1986) and radiosity (Goral et al., 1984) produce faithful imagery given clean meshes and BSDFs, they are computationally prohibitive. Per-scene neural representations (Mildenhall et al., 2020; Kerbl et al., 2023; Hadadan et al., 2021) avoid Monte Carlo tracing but bake each scene into weights, demanding per-scene optimization that cannot generalize to unseen environments. This motivates a generalizable, feed-forward neural rendering approach (Zhang et al., 2026). Recent feed-forward renderers (Zeng et al., 2025; 2026) remain strictly mesh-dependent: by tokenizing at the triangle level ($O(T)$), their dense attention scales quadratically ($O(T^2)$), causing latency and memory to balloon as scene complexity grows (Sec. 4.4). Even concurrent object-level formulations such as RenderFormer++ (Du et al., 2026) remain tied to explicit meshes and must tokenize triangles prior to pooling, while neural light transport embeddings (Xu et al., 2026) predict continuous 3D shading fields via ray queries rather than synthesizing images directly. In contrast, point clouds offer a natural, topology-free representation that interfaces directly with 3D sensors and generative pipelines. Efficiently modeling high-fidelity global illumination from mesh-free representations remains an open challenge.

We present *RADIANT* (**RAD**iance-based **I**nter-**A**ctor **N**eural **T**ransformer), an object-centric Transformer that learns mesh-free feed-forward global illumination. Our design is guided by three observations. First, macroscopic light transport operates predominantly between surface patches and



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

objects rather than diffusing facet by facet across individual triangles. In polygon meshes, tessellation is an artifact of the choice of representation: a surface receives the same illumination whether stored as 50 or 50,000 triangles, or acquired directly as a point cloud. Whereas prior neural renderers require explicit triangle meshes, inheriting primitive-level scaling bottlenecks and polygon discretization constraints, RADIANT is natively *mesh-free*: its sole geometric input is sampled surface points, which are pooled into a fixed budget of 32 surface-patch tokens per object without requiring polygon connectivity. Inter-object light exchange is then resolved among object tokens ($32N$ tokens, scaling scene attention compute as $\mathcal{O}((32N)^2)$ rather than triangle-level $\mathcal{O}(T^2)$), dramatically reducing attention sequence length while unifying mesh-based and point-cloud scenes under one shared model.

Second, inter-object transport depends fundamentally on 3D spatial arrangement, orientation, and materials. Rather than relying on unconstrained attention or materializing expensive $O(N^2)$ pairwise bias matrices, we introduce a *factored radiometric attention bias* (Sec. 3.4) that injects relative directional alignment and material compatibility directly into query-key inner products inside fused attention. Injected before the rotary position encoding, it is modulated by relative 3D pose, and attention remains equivariant to object permutations. *Invariant form-factor attention* (Sec. 3.5) adds the form factor's own geometric arguments, the distance and facing terms between every pair of patches, as exact inner products of per-token features, exactly invariant to rigid motions and inside fused attention. Third, to cleanly decouple local primary surface visibility from global light transport, a lightweight point splatting of input surface points provides a mesh-free primary G-buffer (hit depth, normal, and material). Cross-attention between camera rays and 3D scene tokens in a shared spherical angular frame is then dedicated exclusively to synthesizing non-local indirect illumination, soft shadows, and color bleeding (Sec. 3.6).

Table 1: **Method comparison.** Generalizability, mesh dependency, and learning efficiency, where *Mesh-Free*: no polygon mesh required. *Zero-Opt.*: zero test-time optimization. *Tess.-Inv.*: attention complexity decoupled from polygon tessellation.

| Method | Mesh-Free | Zero-Opt. | Tess.-Inv. | Learning Efficiency[‡] |
|---|---|---|---|---|
| Neural Radiosity | ✗ | ✗[†] | — | — |
| RenderFormer | ✗ | ✓ | ✗ | 80k steps |
| RenderFormer-V2 | ✗ | ✓ | ✗ | n/r[§] |
| RenderFormer++ | ✗ | ✓ | ✗[*] | — |
| **RADIANT (Ours)** | ✓ | ✓ | ✓ | **16k** ($5\times$) |

[†] Per-scene neural optimization: ∼2–4 h. [‡] Training steps to reach 25 dB tone-mapped PSNR on the 1.03M dataset. [*] Dense triangle tokenization prior to pooling. [§] Did not reach 25 dB within 300k steps (Tab. 2). —: not evaluated under our protocol.

As summarized in Tab. 1, RADIANT offers an efficient, generalizable, and scalable framework for feed-forward global illumination. Compared with previous methods, it operates without triangle connectivity (*Mesh-Free*), generalizes without test-time optimization (*Zero-Opt.*), and decouples attention complexity from mesh tessellation (*Tess.-Inv.*). Under a strictly matched protocol (≈206M parameters, identical 1M-scene dataset, loss, and schedule), RADIANT achieves higher sample efficiency and rendering accuracy: it leads RenderFormer by 3.0 dB and RenderFormer-V2 by 7.2 dB tone-mapped PSNR at 300k steps (reaching 30.6 dB), and achieves 25 dB in $5\times$ fewer optimizer steps (16k vs. 80k). Because token count scales linearly with objects rather than primitives ($32N$ tokens vs. $T$ triangles), RADIANT avoids quadratic primitive scaling: while sparse scenes ($\leq 10$ objects) are fast for both RADIANT and RenderFormer ($\leq 70$ ms on an H100), RADIANT scales efficiently to dense scenes (up to 500 objects and 32k triangles per object), maintaining sub-second rendering where triangle transformers stall (Sec. 4.4).

We summarize our contributions below:

- We introduce RADIANT, an efficient and scalable mesh-free Transformer for feed-forward global illumination that operates directly on sampled surface points and generalizes to unseen scenes, decoupling sequence length and attention compute from surface primitives ($\mathcal{O}((32N)^2)$ vs. $\mathcal{O}(T^2)$).
- We propose an object-centric formulation with a factored radiometric attention bias for directional and material transport, an invariant form-factor extension providing geometric inductive biases for patch exchange inside fused attention, and a rasterizer-guided view decoder whose per-ray hit channels restore pixel detail without face connectivity.
- We validate RADIANT on a 1.03M-scene corpus under a strictly matched protocol, demonstrating superior fidelity across all matched budgets, $5\times$ higher learning efficiency during

training, and scalable inference that avoids quadratic primitive bottlenecks on complex scenes. All code, trained models, and data will be publicly released.

## 2 Related Work

**Per-scene neural rendering.** Early work on neural rendering focused on overfitting models to individual scenes. Neural radiance fields (Mildenhall et al., 2020; Müller et al., 2022), 3D Gaussian splatting (Kerbl et al., 2023), and Neural Radiosity (Hadadan et al., 2021), which directly minimizes the residual of the rendering equation (Kajiya, 1986), use per-scene optimization to encode a scene, and require re-optimization for any other. In contrast, we focus on feed-forward rendering of unseen scenes from explicit descriptions without test-time optimization.

**Feed-forward rendering of explicit scenes.** RenderFormer (Zeng et al., 2025) established that a transformer (Vaswani et al., 2017) can act as a general renderer: triangles with reflectance attributes become tokens, a view-independent stage models triangle-to-triangle transport, and a view-dependent stage decodes ray bundles to pixels. RenderFormer-V2 (Zeng et al., 2026) introduces windowed attention and summary tokens for participating media and environment lighting, but remains fundamentally centered on triangle primitives. Concurrent RenderFormer++ (Du et al., 2026) similarly pools triangle tokens into object-level representations with a G-buffer decoder. However, it requires explicit triangle-mesh input and encodes triangle tokens before object-level aggregation ($O(T)$). Lacking public code or weights, empirical benchmarking is precluded; analytically, its initial $O(T_{\text{in}})$ tokenization contrasts with RADIANT's natively mesh-free PointNet++ patch tokens (Qi et al., 2017b). TransGI (Deng et al., 2025) targets dynamic object-centric radiance transfer via light probes and material MLPs rather than feed-forward multi-object 3D transport. While instance-structured 3D tokenization has emerged in reconstruction (Yoo et al., 2026), we show object-centric tokens are uniquely suited to global illumination, where macroscopic transport couples surfaces rather than individual facets. Earlier non-mesh methods overfit small scene sets (Sanzenbacher et al., 2020), compose proxies without full geometric interreflections (Granskog et al., 2021), or embed isolated shapes for generation (Zhang et al., 2023). RADIANT provides an end-to-end feed-forward framework that directly renders full HDR images from unorganized multi-object point clouds, bypassing polygon tessellation entirely.

**Alternative representations for neural global illumination.** Prior neural methods explore varied illumination representations: Xu et al. (Xu et al., 2026) predict indirect transport by querying continuous neural fields over point clouds, LightFormer (Ren et al., 2024) aggregates virtual point lights with pixel–light attention, and Deep Shading (Nalbach et al., 2017) synthesizes illumination from rasterized G-buffers via CNNs. In contrast, RADIANT represents scenes as object-centric surface tokens, resolving full 3D transport in attention while using primary rasterization to ground camera ray visibility.

**Rotary embeddings and attention biases.** Relative 3D spatial arrangements govern radiative exchange between surfaces. Rotary position embeddings (RoPE) (Su et al., 2024) encode relative coordinate offsets directly into query-key inner products, maintaining compatibility with fused kernels (Dao et al., 2022). PRoPE (Li et al., 2025) treats camera frustums as relative encodings for view synthesis, and we adopt its CamRay ray parameterization. While standard additive pairwise bias matrices incur prohibitive $O(N^2)$ memory, low-rank factorizations such as FlashBias (Wu et al., 2025) avoid this cost by decomposing biases into query-key inner products. Our factored radiometric bias injects material and physical area proxies *prior* to the rotary rotation, directly modulating geometric interactions by relative 3D pose inside fused attention. We also incorporate FiLM (Perez et al., 2018) and register tokens (Darcet et al., 2024) for feature-wise conditioning and scratch capacity.

## 3 Method

### 3.1 Problem setup

A scene is defined as a set of $N$ objects $\{\mathcal{O}_i\}_{i=1}^{N}$, each given by surface geometry with spatially varying material attributes (diffuse albedo, specular reflectance, roughness, and emission; emissive objects are the light sources), together with a query camera with intrinsics and view direction $(K, [R \mid t])$. The task is to produce, in one feed-forward pass and for scenes never seen in training,

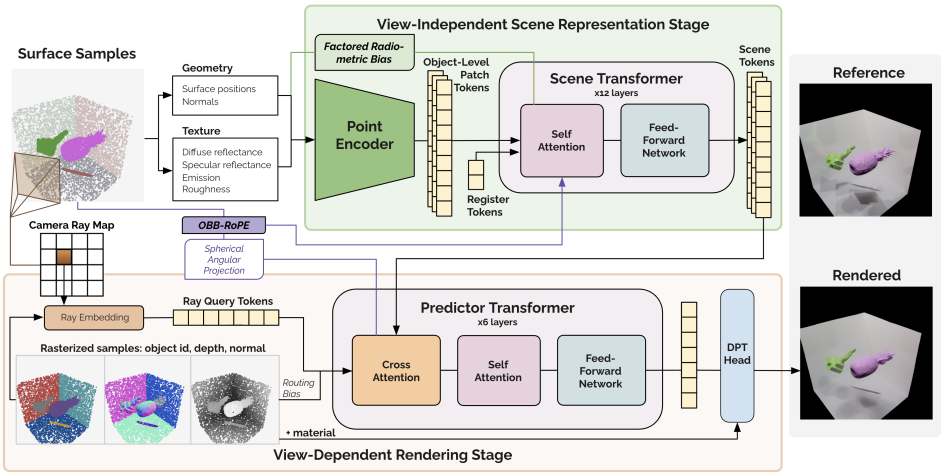



Figure 1: **Overview of RADIANT**'s two-stage rendering pipeline. In the view-independent first stage, discrete surface samples are pooled into 32 patch tokens per object and a scene transformer resolves 3D light transport with OBB-RoPE and factored radiometric attention bias. In the view-dependent second stage, for each query viewpoint, the predictor transformer cross-attends camera rays to scene tokens guided by rasterized geometry channels, spherical RoPE, and object routing to decode the HDR image.

the HDR radiance image $L \in \mathbb{R}^{H\times W\times 3}$ that a path tracer would render. We represent each object by $P{=}2{,}048$ surface samples, each carrying a 16-dimensional feature: position, normal, and the ten material parameters at the sample. RADIANT is natively *mesh-free*: a mesh-based object undergoes uniform surface sampling, and a point cloud with normals and materials is provided directly as model input without surface reconstruction. All our experiments use mesh-derived samples.

### 3.2 Architecture

Fig. 1 illustrates our two-stage pipeline. First, a point encoder maps each object's samples to $M{=}32$ surface-patch tokens (Sec. 3.3), and global transport is modeled by a *scene transformer* (12 self-attention layers, width 1,024, 8 heads, $4\times$ MLP, 97.0M parameters) using OBB-RoPE (Sec. 3.3) and factored radiometric attention bias (Sec. 3.4). In the second stage, a *predictor* (6 alternating cross/self-attention layers, 96.1M parameters) decodes views by cross-attending ray tokens (built from CamRay and point-splat rasterization; Sec. 3.6) against fused scene features. Finally, a DPT head (Ranftl et al., 2021) upsamples the $4 \times 4$ patch tokens to HDR radiance, decoupling compute from polygon tessellation.

### 3.3 Object Tokenization

A PointNet-style encoder (Qi et al., 2017a) (10.0M parameters) embeds $P$ surface samples per object from a canonicalized frame. Using PointNet++ set abstraction (Qi et al., 2017b), farthest-point sampling selects $M{=}32$ centroids per object with 128 nearest neighbors, where max-pooling and linear projection yield 32 patch tokens of width 1,024 per object. Sixteen learnable register tokens (Darcet et al., 2024) absorb extraneous attention weight ($\leq 0.05$ dB ablation effect; Sec. 4.3). By construction, the scene transformer is equivariant to object set permutations, ensuring that rendered imagery is invariant to object input ordering (Prop. 1, Appendix B).

To express relative 3D geometry, we inject object pose into attention queries and keys via an object-centric rotary embedding (*OBB-RoPE*). Each token is assigned a 12-dimensional oriented bounding box descriptor $\mathbf{x}^{(i)} \in \mathbb{R}^{12}$, concatenating its 3D centroid with its three principal axes scaled by their extents (computed in world space for the scene transformer and camera space for the predictor; Appendix F). This rotation composes into a relative spatial attention function $R(\mathbf{x}^{(j)} - \mathbf{x}^{(i)})$ (Prop. 2).

### 3.4 Factored Radiometric Attention Bias

Diffuse light transport between objects is strongly governed by their relative distance, orientation, and material properties, traditionally formalized by the radiative *form factor* (or configuration factor)

(Goral et al., 1984). Standard vision transformers lack this domain-specific inductive bias. While an explicit $N \times N$ attention bias matrix $\phi_{ij}$ could introduce this prior, it incurs $\mathcal{O}(N^2)$ memory overhead and breaks compatibility with fused kernels like FlashAttention (Dao et al., 2022). To preserve linear $\mathcal{O}(N)$ memory scaling for token representations and retain fused-kernel execution, we introduce a low-rank, factored radiometric attention bias. Instead of an explicit pairwise matrix, we encode per-object *receiver* features into queries and *sender* features into keys $\mathbf{q}'_i = \mathbf{q}_i + \delta\mathbf{q}_i$ and $\mathbf{k}'_j = \mathbf{k}_j + \delta\mathbf{k}_j$, where $\delta\mathbf{q}_i$ and $\delta\mathbf{k}_j$ are learned encodings of eight radiosity-relevant per-object statistics shared by the object's tokens: a log area proxy, mean diffuse albedo, mean emission, and mean specular reflectance. The logit expands as

$$\mathbf{q}_i'^{\top}\mathbf{k}'_j = (\mathbf{q}_i + \delta\mathbf{q}_i)^{\top}(\mathbf{k}_j + \delta\mathbf{k}_j) = \underbrace{\mathbf{q}_i^{\top}\mathbf{k}_j}_{\text{content}} + \underbrace{\delta\mathbf{q}_i^{\top}\delta\mathbf{k}_j}_{\approx \phi_{\mathrm{GI}}} + \underbrace{\mathbf{q}_i^{\top}\delta\mathbf{k}_j + \delta\mathbf{q}_i^{\top}\mathbf{k}_j}_{\text{cross terms}}, \tag{1}$$

where standard semantic attention is augmented by a pure geometric compatibility term $\delta\mathbf{q}_i^{\top}\delta\mathbf{k}_j$ and adaptive content-geometry cross-terms. In our implementation, QK-normalization ($\hat{\mathbf{q}}'_i = \mathbf{q}'_i/\|\mathbf{q}'_i\|_2$, $\hat{\mathbf{k}}'_j = \mathbf{k}'_j/\|\mathbf{k}'_j\|_2$) is applied post-injection to stabilize logit magnitude, while preserving relative directional and material alignment rankings. Note that while a fixed budget of 32 patch tokens per object allocates uniform sequence capacity regardless of physical dimensions, the sender key modification $\delta\mathbf{k}_j$ incorporates the object's log surface area proxy, allowing attention weights to scale with physical radiative emission. We inject $\delta\mathbf{q}$ and $\delta\mathbf{k}$ *before* applying the rotary position embedding OBB-RoPE. Because the rotary transformation $R(\cdot)$ acts on the augmented vector sum, the pure geometric term implicitly expands to

$$\delta\mathbf{q}_i^{\top} R(\mathbf{x}^{(j)} - \mathbf{x}^{(i)})\delta\mathbf{k}_j, \tag{2}$$

scoring receiver–sender material compatibility modulated by relative spatial distance and orientation. Injecting these features *after* rotation would collapse this interaction into a pose-independent inner product $\delta\mathbf{q}_i^{\top}\delta\mathbf{k}_j$ (Prop. 3). Non-separable transport effects (e.g., occlusion) are naturally handled by the full attention mechanism. Finally, two FiLM pathways (Perez et al., 2018) (2.1M parameters each) further condition each object's tokens on rotation-invariant sample shape statistics and, in the view-dependent stage, its pose relative to the query camera.

### 3.5 Invariant Form-Factor Attention

While rotary position embeddings capture relative positions, they do not directly expose the fundamental geometric invariants that govern diffuse light exchange, namely, relative distance $r_{ij}$ and the angles $(\theta_i, \theta_j)$ between surface normals and the connecting displacement vector $\mathbf{\Delta}_{ij} = \mathbf{c}_j - \mathbf{c}_i$. To address this, we formulate an exact inner-product factorization that embeds $O(3)$-invariant geometric quantities directly into the query-key dot product. Each token carries its centroid $\mathbf{c}_i$ and unnormalized mean normal $\mathbf{m}_i$ (computed over its 128 surface samples). We construct a 4-dimensional geometric invariant vector $\mathbf{G}_{ij} \in \mathbb{R}^4$ consisting of four physical terms (embedded via a 16-dimensional factored query-key representation):

$$\mathbf{G}_{ij} = \Big( \underbrace{-\|\mathbf{\Delta}_{ij}\|^2}_{\text{distance kernel}}, \quad \underbrace{\mathbf{m}_i^{\top}\mathbf{\Delta}_{ij}}_{\text{receiver facing}}, \quad \underbrace{-\mathbf{m}_j^{\top}\mathbf{\Delta}_{ij}}_{\text{sender facing}}, \quad \underbrace{\mathbf{m}_i^{\top}\mathbf{m}_j}_{\text{mutual orientation}} \Big). \tag{3}$$

Observe that each of these four invariant scalar terms can be expressed as an exact inner product between receiver-side and sender-side features, e.g., squared distance expands as $-\|\mathbf{\Delta}_{ij}\|^2 = \langle [2\mathbf{c}_i, -\|\mathbf{c}_i\|^2, -1], [\mathbf{c}_j, 1, \|\mathbf{c}_j\|^2] \rangle$. By appending these factored features directly to the scene transformer's queries and keys, we inject a learned form-factor prior without explicit $N \times N$ matrices. Per layer and attention head, these terms are modulated by receiver and sender gains $a_t(i)$ and $b_t(j)$ predicted from the radiosity statistics (Sec. 3.4). This adds a learned scalar bias $\sum_{t=1}^{4} a_t(i) b_t(j) G_{ij,t}$ directly to each attention logit inside the fused attention kernel, evaluating a Gaussian distance decay with material-dependent orientation gating.

**Theoretical Properties & Scope.** These terms provide the exact $O(3)$-invariant quadratic arguments of the differential form factor $dF_{ij}$ (Prop. 5). Non-quadratic physical properties (namely, non-negative cosine clamping $[\cos\theta]_+$, $1/r_{ij}^2$ distance falloff, and visibility occlusion) are non-finite-rank and are modeled by subsequent transformer layers. (Note: main $512 \times 512$ experiments in Sec. 4.2 evaluate the base RADIANT architecture; this invariant form-factor augmentation is ablated in Sec. 4.3).

### 3.6 View-Dependent Decoding: Ray Tokens, Rasterizer Channels, and Spherical Cross-Attention

**Ray tokens.** We adopt an unnormalized variant of the CamRay parameterization (Li et al., 2025): each pixel's *CamRay*, proportional to $K^{-1}[u, v, 1]^\top$ in an OpenGL camera frame, preserves the linearity of the image plane. Pixels are grouped into $4 \times 4$ patches. The patch CamRay map, together with the per-pixel channels described below, is linearly projected and layer-normalized into a single ray token positioned at the patch's CamRay centroid.

**Rasterizer channels and object routing.** Rather than forcing cross-attention to resolve primary ray intersection from scratch, we decouple local primary surface visibility from global light transport. A depth-tested point-splat rasterization of surface samples for the query camera ($7 \times 7$ dilation) provides a mesh-free primary G-buffer with visible object identity, depth, normal, and material per pixel. Because this buffer contains only primary surface hits without indirect illumination or shadows, all multi-bounce transport is synthesized exclusively through cross-attention to the 3D scene tokens. The resulting log-depth, camera-space normal, and hit mask are patchified alongside the CamRay map, ensuring each ray token carries its primary hit geometry. Object identities enter cross-attention as a *routing bias*: the attention logit between a ray patch and an object's tokens receives an additive bias proportional to the fraction of patch pixels rasterized to that object, scaled by a learned per-layer signed gate (initialized to zero). For output generation, rasterized normal, materials, and depth (14 channels) are added to full-resolution DPT features via a zero-initialized $1 \times 1$ convolution.

**Spherical cross-attention RoPE.** Cross-attention requires comparing 3D object positions against 2D image-plane coordinates under a unified relative encoding. We project both into a bounded spherical angular space with range $[-\pi, \pi]$: objects at camera-space position $(X, Y, Z)$ map to $(\mathrm{atan2}(X, -Z),\ \mathrm{atan2}(Y, -Z),\ -Z)$ and rays with plane coordinates $(u, v)$ to $(\arctan u,\ \arctan v,\ \bar{d})$, where $\bar{d}$ is the mean rasterized hit depth of the patch (or 0 for un-intersected rays). Under RoPE subtraction, the difference in the first two coordinates directly measures the angular offset between the ray and the object from the camera's perspective, while the third coordinate provides explicit depth information for occlusion reasoning. These angular coordinates coincide exactly when a ray points toward the object's centroid (Prop. 4).

### 3.7 Training objective

Our model is trained end-to-end. The network predicts log radiance, $\hat{y} = \log(1 + L)$, and following Zeng et al. (2025), the loss $\mathcal{L}$ is the mean absolute error in that domain plus a weighted perceptual term (Zhang et al., 2018) on the tone-mapped images $\tau(L) = \mathrm{clip}_{[0,1]}(\log_2(1 + L))$:

$$\mathcal{L} = \left\| \hat{y} - \log(1 + L^\star) \right\|_1 + 0.05\,\mathrm{LPIPS}\big(\tau(\hat{L}), \tau(L^\star)\big), \tag{4}$$

where $L^\star$ is ground-truth radiance and $\hat{L} = \exp(\min(\hat{y}, 10)) - 1$. We optimize with AdamW using peak learning rate $5 \times 10^{-5}$, 8,000 warm-up steps, cosine decay to $10^{-6}$, and global batch size 24. All models share this loss, tone map, LPIPS backbone, and optimizer setup (Appendix C).

## 4 Experiments

### 4.1 Experimental setup

**Data.** Following the RenderFormer data generation protocol (Zeng et al., 2025), we construct procedural scenes of one to six Objaverse objects (Deitke et al., 2023) from a curated pool of 18,085 meshes in a template room with randomized wall pruning, area lights (1–8 emissive entities per scene, mean 4.5; Tab. 7), and per-shading-group materials. Four ground-truth views per scene are rendered via Blender Cycles path tracing at $512^2$ with 512 samples per pixel (Appendix D). To ensure a controlled comparison, all models are trained from scratch on our 1.03M-scene dataset under identical data distributions, without relying on external pre-rendered datasets or pretrained weights. The last 0.1% of scenes form the held-out test split (1,028 scenes, 4,112 images which sample from the same mesh pool as training but feature strictly novel spatial arrangements, lighting, materials, and camera viewpoints). Each model reads identical scenes in its native geometric format (surface points for RADIANT, triangle meshes for RenderFormer baselines).

Table 2: Held-out rendering quality under the matched-budget protocol ($512^2$, 4,112 images). All three models at matched step 300k. Bold indicates best performance.

| Model | Step | PSNR ↑ (per-image / pooled) | LPIPS ↓ | $\ell_1(\log)$ ↓ |
|---|---|---|---|---|
| RADIANT (ours) | 300k | **30.62 / 26.99** | **0.053** | **0.015** |
| RenderFormer | 300k | 27.62 / 24.99 | 0.093 | 0.023 |
| RenderFormer-V2 | 300k | 23.43 / 22.06 | 0.142 | 0.051 |

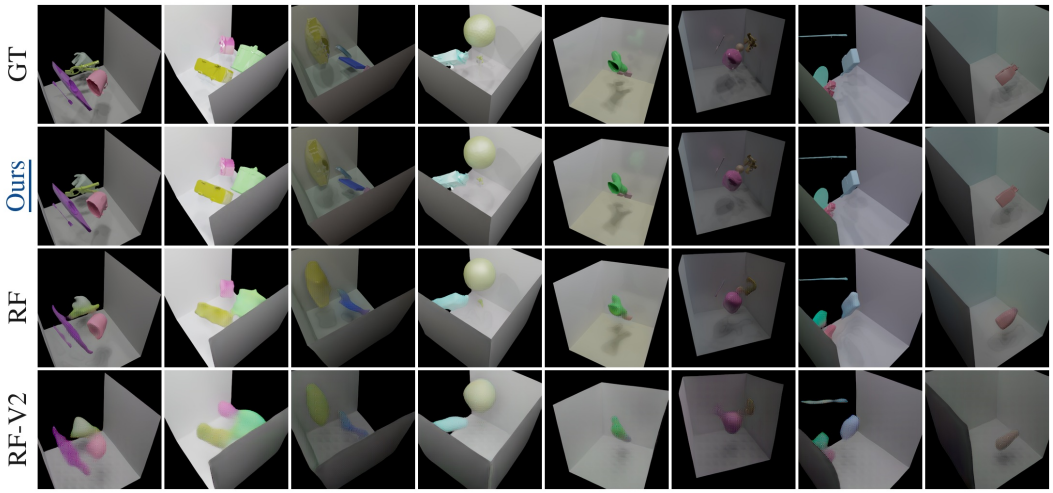



Figure 2: Qualitative comparisons on crowded held-out scenes at matched step 300k. Rows: Ground Truth (GT), RADIANT (Ours), RenderFormer (RF), and RenderFormer-V2 (RF-V2).

**Matched-budget protocol.** We compare three models of matched size: RADIANT (205.9M parameters), RenderFormer Base (205.2M) (Zeng et al., 2025), and RenderFormer-V2 in its 200M configuration (207.6M; Appendix C). All models are trained from scratch at $512^2$ on an $8\times$ NVIDIA A100 node with a global batch size of 24 (3 scenes per device, without data augmentation) following the recipe in Sec. 3.7, using gradient clipping at 1.0, AdamW optimizer, bfloat16 mixed precision with FP32 master weights, and identical loss formulations. Training follows a 200k-step cosine schedule, continued from step 200k on the tail of a 300k-step cosine schedule; the three models are compared at matched step 300k, with checkpoints evaluated throughout training (Sec. 4.2). Main checkpoints are trained with seed 42. Across multiple seeds at 25k–100k steps, performance variance is minimal (0.05–0.08 dB spread), and paired bootstrap 95% confidence intervals confirm that our $+3.0$ dB and $+7.2$ dB margins are statistically significant (Appendix G).

**Metrics.** We evaluate all models on the same 4,112 held-out image set using three metrics: (1) PSNR after Reinhard tone mapping (Reinhard et al., 2002) and sRGB conversion, evaluated both *per-image* and *pooled* across all pixels prior to the log, (2) LPIPS (Zhang et al., 2018) using the AlexNet backbone, and (3) the mean absolute error in log radiance (the data term of the loss).

### 4.2 Rendering Quality

RADIANT renders more accurately than both RenderFormer generations at every budget measured, and reaches their quality with a fraction of their optimizer steps. Tab. 6 (Appendix C) lists every evaluated step. At matched step 6k, RADIANT leads the RenderFormer architecture by $+6.2$ dB per-image and halves its LPIPS. By matched step 300k, the end of the extended schedule (Sec. 4.1), it leads RenderFormer by $+3.0$ dB per-image ($+2.0$ dB pooled) and RenderFormer-V2 by $+7.2$ dB (30.62 versus 27.62 and 23.43 dB), with 57% of RenderFormer's LPIPS and 37% of RenderFormer-V2's (0.053 versus 0.093 and 0.142). At intermediate 50k and 100k checkpoints, the leads were $+4.8$ and $+8.0$ dB and $+3.8$ and $+7.2$ dB, respectively, while RADIANT at 16k already exceeds RenderFormer at 50k (25.15 versus 23.25 dB). RADIANT maintains a consistent lead across scene complexity: 2.8–3.4 dB in every entity-count bucket (Tab. 4). Selected over all evaluated checkpoints on the same split, identically

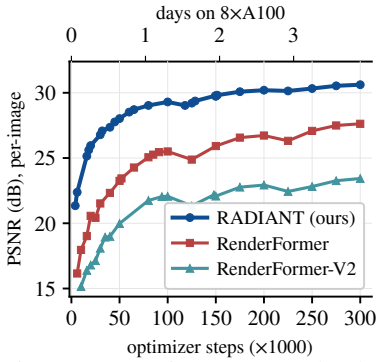



Figure 3: Learning curves under the matched protocol on 4,112 held-out images. Top axis: training days on an $8\times$A100 node (1.12 s/step). RADIANT converges substantially faster and reaches higher quality.

Table 3: Ablation of key mechanisms ($256^2$, 100k steps, seed 42). PSNR per-image / pooled; $\Delta$: difference from full model. Extended ablations in Tab. 16.

| Variant | PSNR ↑ | LPIPS ↓ | $\ell_1(\log)$ ↓ | $\Delta$ |
|---|---|---|---|---|
| Full model | 29.36 / 25.88 | 0.063 | 0.021 | — |
| No factored radiometric bias (§3.4) | 26.92 / 24.59 | 0.072 | 0.042 | $-2.43$ |
| No rasterizer channels (hit geometry, route bias) | 26.21 / 24.03 | 0.099 | 0.027 | $-3.14$ |
| + Invariant form-factor attention (§3.5) | **29.97 / 26.15** | **0.058** | **0.019** | $+0.61$ |
| + float32 scene attention only (precision control) | 29.43 / 25.91 | 0.063 | 0.021 | $+0.07$ |
| Invariant attention in place of OBB-RoPE | 29.83 / 26.11 | 0.060 | **0.019** | $+0.47$ |

for the three models (Appendix C), the best checkpoints are RADIANT at 300k (30.62 dB), RenderFormer at 294k (27.63 dB), and RenderFormer-V2 at 300k (23.43 dB).

On crowded held-out scenes at matched step 300k (Fig. 2), RADIANT resolves fine geometric structures, sharp contact shadows, albedo textures, and inter-reflections. In contrast, both RenderFormer baselines exhibit geometric blurring, color shifts, and patch-grid artifacts. Further held-out scenes appear in Appendix A. Across all evaluated steps, RADIANT consistently outperforms both baselines, matching or exceeding RenderFormer's 100k-step quality by step 20k and RenderFormer-V2's by step 6k (Fig. 3), using five to sixteen times fewer optimizer steps.

Beyond complete synthetic geometry, mesh-free representations naturally accommodate incomplete or noisy geometry. Without fine-tuning, RADIANT reaches 27.50 dB on single-view depth scans missing 52–72% of surface points, against 32.43 dB from the complete samples (Tab. 22). With 400 steps of per-scene fine-tuning on the four views of each scene, PSNR on these views reaches 36.50 dB, within 0.74 dB of the same fine-tuning from complete geometry (Fig. 4, Appendix I).

### 4.3 Ablations

As reported in Tab. 3, ablations on 2,948 held-out images isolate the physical and architectural drivers of rendering fidelity under a controlled 100k-step schedule at $256^2$:

**3D light transport modeling.** Holding the rasterizer and view decoder strictly fixed, omitting the factored radiometric bias in the scene transformer degrades PSNR by $-2.43$ dB ($-2.47 \pm 0.05$ dB across three random seeds; Tab. 18) and doubles log-radiance error from $0.021$ to $0.042$. As training progresses from 25k to 100k steps, this performance gap widens from $-1.32$ dB to $-2.43$ dB (Tab. 19), indicating that the network increasingly relies on relative directional alignment and material compatibility to resolve multi-bounce inter-reflections. Incorporating invariant form-factor attention (Sec. 3.5) boosts PSNR by $+0.61$ dB ($+0.52 \pm 0.08$ dB across seeds) and reduces LPIPS by 7%. Our float32 precision control confirms that numerical precision accounts for only $+0.07$ dB of this gain, while substituting invariant attention directly for OBB-RoPE yields a $+0.47$ dB improvement (29.83 dB). Thus, physically grounded geometric inductive biases provide over 3 dB of cumulative fidelity gains inside fused attention.

**Explicit visibility grounding.** Removing rasterizer hit channels degrades PSNR by $-3.14$ dB ($3.12 \pm 0.05$ dB across seeds) and increases LPIPS by 57%, despite the primary G-buffer containing zero multi-bounce lighting on its own (Tab. 21). Conversely, local rasterized G-buffer decoding without 3D scene tokens causes severe degradation (Appendix H), confirming that rasterization grounds local primary visibility while cross-attention resolves global non-local transport.

**Decoder capacity and token efficiency.** At matched $4{,}096$ ray tokens ($256^2$ vs. RenderFormer at $512^2$), RADIANT leads by $> 2.6$ dB, demonstrating that performance gains do not depend on decoder token density. Auxiliary components (FiLM, OBB-RoPE, registers) contribute $\leq 0.2$ dB (Tab. 16). Furthermore, varying the patch token budget eightfold ($M \in \{8, 16, 32, 64\}$) alters PSNR by $\leq 0.12$ dB (Tab. 12): allocating 16 patch tokens achieves 29.47 dB (matching 32 tokens at 29.36 dB) while reducing training latency by 17%, establishing that compact patch representations reliably capture macroscopic surface transport.

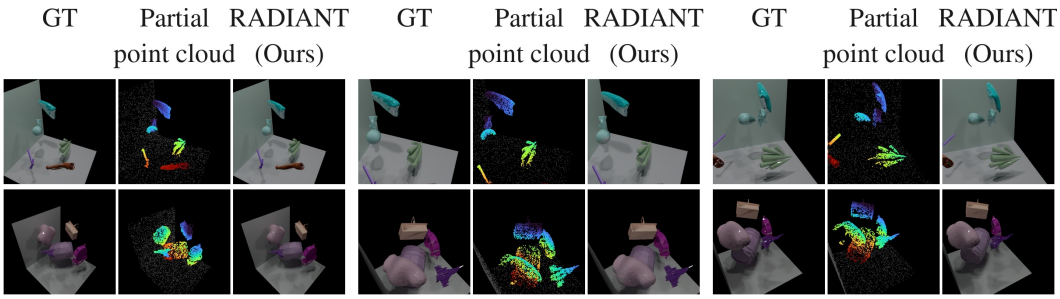



Figure 4: Rendering from partial point clouds (Appendix I). Two held-out scenes (rows), three views each (scanning camera excluded). Left: ground truth reference. Middle: partial point cloud from single-view depth scans per object (unscanned surfaces absent). Right: RADIANT rendered after 400 steps of per-scene fine-tuning.

### 4.4 Inference efficiency and scaling with scene complexity

Because token count in RADIANT scales linearly with object count ($32N$) rather than primitives ($T_{\text{in}}$), sequence length is decoupled from tessellation: while scene self-attention is $\mathcal{O}((32N)^2)$ in token count, it is independent of triangle count. Inference latency scales favorably: at $512^2$, RADIANT slows by only $5.5\times$ from 1 to 500 objects, whereas RenderFormer slows by $944\times$ ($231\times$ with FlashAttention). In small benchmark scenes (median 10 entities, $\sim$3.3k triangles), RenderFormer is faster (35 vs. 69 ms) because RADIANT incurs fixed overhead from 16,384 ray tokens, DPT decoding, and rasterization, before quadratic triangle costs dominate (Tab. 8).

Beyond $\approx 20$ objects or $\approx 2{,}048$ triangles/object (Tab. 9), RADIANT leads decisively: achieving up to $50.7\times$ speedup at 500 objects ($19.3\times$ vs. FlashAttention) and $431\times$ at 32k triangles/object ($163\times$ vs. FlashAttention). Fig. 6 and Appendix E provide detailed scaling curves and latency breakdowns independent of tessellation. Within the training regime (1–18 entities), RADIANT maintains its $> 3$ dB advantage (Tab. 4).

**Batched throughput and multi-view amortization.** On small scenes, batched execution favors RADIANT at $128^2$ and $256^2$ for batch sizes $\geq 8$ (Tab. 13), achieving up to 4.5 ms and 10.9 ms per frame ($2.8\times$ and $1.3\times$ faster than RenderFormer with FlashAttention). Furthermore, in multi-view rendering where the view-independent scene representation is computed once and replayed across viewpoints (Tab. 14), RADIANT's scene transformer requires only 30.5 ms for 50 objects compared to 117.6 ms for RenderFormer under FlashAttention (227 ms under standard SDPA). At $256^2$, amortized per-view latency at 64 views favors RADIANT (16.1 vs. 22.4 ms per view), demonstrating practical efficiency for multi-camera rendering and dynamic scene navigation.

## 5 Conclusion

We presented RADIANT, a natively mesh-free, object-centric feed-forward transformer for global illumination: objects enter as surface samples pooled into 32 patch tokens, transport is resolved in 3D without triangles, and pixel detail is restored per view. Under a matched protocol, RADIANT outperforms both RenderFormer generations across training horizons with superior learning efficiency ($5\times$–$16\times$ fewer steps), maintaining efficient scaling on dense scenes where triangle renderers encounter quadratic bottlenecks. By shifting from triangle-level discretization to patch-level light transport, RADIANT shows that generalizable feed-forward global illumination is achievable directly from mesh-free surface representations.

**Limitations and future work.** A fixed budget of 32 patch tokens per object over-allocates tokens to simple shapes while under-allocating to intricate ones, motivating adaptive token allocation based on geometric complexity. Because ray cross-attention scales quadratically with image resolution, multi-scale or windowed cross-attention will be valuable at $1024^2$ and beyond. Our rasterizer decoding assumes direct visibility, leaving participating media, refraction, and distant environment lighting to future work. In addition, our latency scaling up to 500 objects evaluates systems throughput; assessing rendering fidelity at this extreme scale is currently constrained by the difficulty of generating path-traced ground truth. While our formulation handles per-sample material attributes, incorporating high-frequency UV texture maps and extending the architecture to unsegmented sensor scans remain open directions. Finally, the modular object-centric structure provides natural interfaces for inverse graphics, modular uncertainty attribution, and integration with physics simulators.

### AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

Large language models (LLMs) and AI assistants were used for polishing manuscript prose, assisting with figure visualization and plotting scripts, and executing and monitoring experimental runs. All core scientific ideas, architectural designs, mathematical formulations, and experimental protocols originated entirely from the human authors. All experimental execution logs and reported results were systematically inspected, verified, and accepted by the human authors, with no generative AI used to fabricate, impute, or extrapolate findings. The authors take full responsibility for the final content of this paper.

### ETHICS STATEMENT

This work uses only synthetic scenes assembled from publicly available Objaverse assets (Deitke et al., 2023) and involves no human subjects or personal data. All 3D assets are used in compliance with their respective Creative Commons licenses (CC-BY 4.0, CC-BY-SA, and CC0), with individual artist attribution and licenses preserved in our released scene manifest.

### REPRODUCIBILITY STATEMENT

The matched-budget protocol (Sec. 4.1) and Appendices C–F specify model architectures, parameter counts, dataset generation, loss functions, optimizer configurations, and evaluation schedules for all models. All models are trained under a strictly matched configuration and evaluated using a unified pipeline. We will release our complete data-generation pipeline, scene corpus, training configurations, and evaluation code. Anonymized code, configurations, and checkpoint evaluation scripts are provided in the supplementary material and will be made publicly available upon publication.

## REFERENCES

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3D objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Yijie Deng, Lei Han, and Lu Fang. TransGI: Real-time dynamic global illumination with object-centric neural transfer model. *IEEE Transactions on Visualization and Computer Graphics*, 2025.

Huangsheng Du, Haoran Zhu, Youcheng Cai, Jingyang Meng, and Ligang Liu. Renderformer++: Scalable and physics-informed feed-forward neural rendering. *arXiv preprint arXiv:2606.30380*, 2026. v2, August 2026.

Cindy M. Goral, Kenneth E. Torrance, Donald P. Greenberg, and Bennett Battaile. Modeling the interaction of light between diffuse surfaces. In *Proceedings of SIGGRAPH*, 1984.

Jonathan Granskog, Till N. Schnabel, Fabrice Rousselle, and Jan Novák. Neural scene graph rendering. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 2021.

Saeed Hadadan, Shuhong Chen, and Matthias Zwicker. Neural radiosity. *ACM Transactions on Graphics*, 40(6), 2021.

James T. Kajiya. The rendering equation. In *Proceedings of SIGGRAPH*, 1986.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2507.10496.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4), 2022.

Oliver Nalbach, Elena Arabadzhiyska, Dushyant Mehta, Hans-Peter Seidel, and Tobias Ritschel. Deep shading: Convolutional neural networks for screen space shading. *Computer Graphics Forum*, 36(4):65–78, 2017. doi: 10.1111/cgf.13225.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017a.

Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017b. arXiv:1706.02413.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. *ACM Transactions on Graphics*, 21(3):267–276, 2002.

Haocheng Ren et al. LightFormer: Light-oriented global neural rendering in dynamic scene. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 2024.

Paul Sanzenbacher, Lars Mescheder, and Andreas Geiger. Learning neural light transport. *arXiv preprint arXiv:2006.03427*, 2020.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Hermann Weyl. *The Classical Groups: Their Invariants and Representations*. Princeton University Press, 1946.

Haixu Wu, Minghao Guo, Yuezhou Ma, Yuanxu Sun, Jianmin Wang, Wojciech Matusik, and Mingsheng Long. FlashBias: Fast computation of attention with bias. *arXiv preprint arXiv:2505.12044*, 2025.

Bing Xu, Mukund Varma T, Cheng Wang, Tzu-Mao Li, Lifan Wu, Bartlomiej Wronski, Ravi Ramamoorthi, and Marco Salvi. A generalizable light transport 3d embedding for global illumination. *ACM SIGGRAPH 2026 Conference Papers*, 2026. arXiv:2510.18189.

Mijin Yoo, In Cho, Subin Jeon, Jiwoo Lee, Eunbyung Park, and Seon Joo Kim. Scenes as objects, not primitives: Instance-structured 3D tokenization from unposed views. *arXiv preprint arXiv:2606.29513*, 2026.

Chong Zeng, Yue Dong, Pieter Peers, Hongzhi Wu, and Xin Tong. Renderformer: Transformer-based neural rendering of triangle meshes with global illumination. In *ACM SIGGRAPH 2025 Conference Papers*, 2025. doi: 10.1145/3721238.3730595.

Chong Zeng, Yue Dong, Pieter Peers, Lvmin Zhang, and Maneesh Agrawala. Renderformer-v2: Neural rendering with heterogeneous scene primitives. In *European Conference on Computer Vision (ECCV)*, 2026. arXiv:2609.05738.

Biao Zhang, Jiapeng Tang, Matthias Nießner, and Peter Wonka. 3DShape2VecSet: A 3D shape representation for neural fields and generative diffusion models. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 2023. arXiv:2301.11445.

Jiahui Zhang, Yuelei Li, Anpei Chen, et al. Advances in feed-forward 3D reconstruction and view synthesis: A survey. *Computer Graphics Forum*, 2026.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.