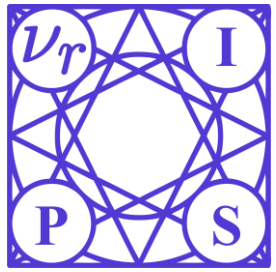
Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations

Vincent Sitzmann Michael Zollhöfer Gordon Wetzstein

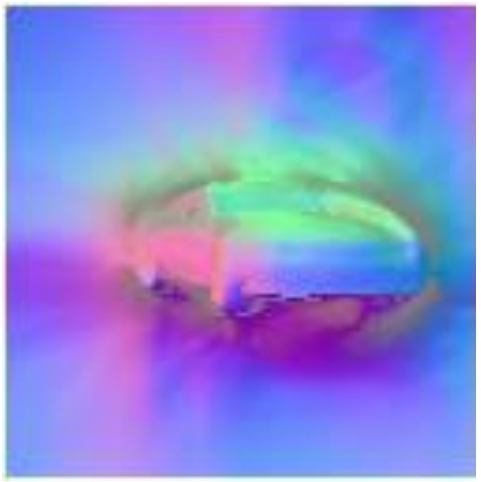


Stanford University

Single image
camera pose
intrinsics

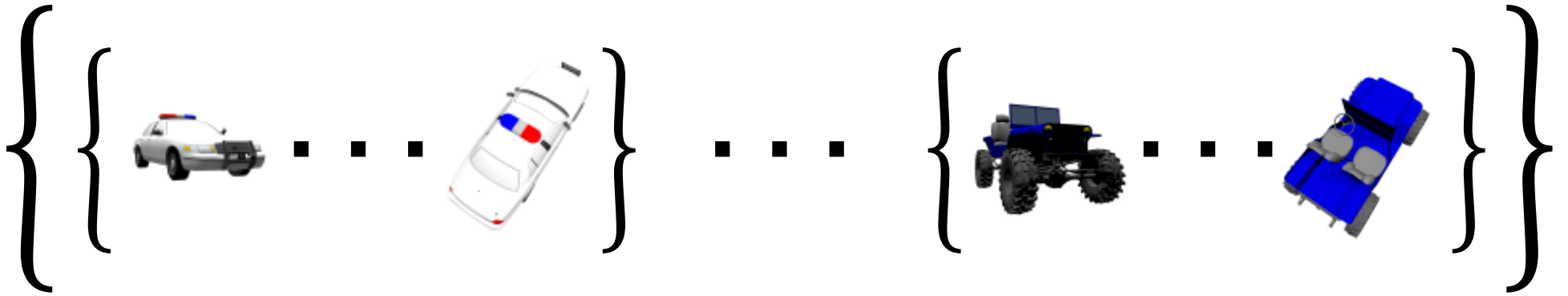
Novel Views

Surface Normals



Self-supervised Scene Representation Learning

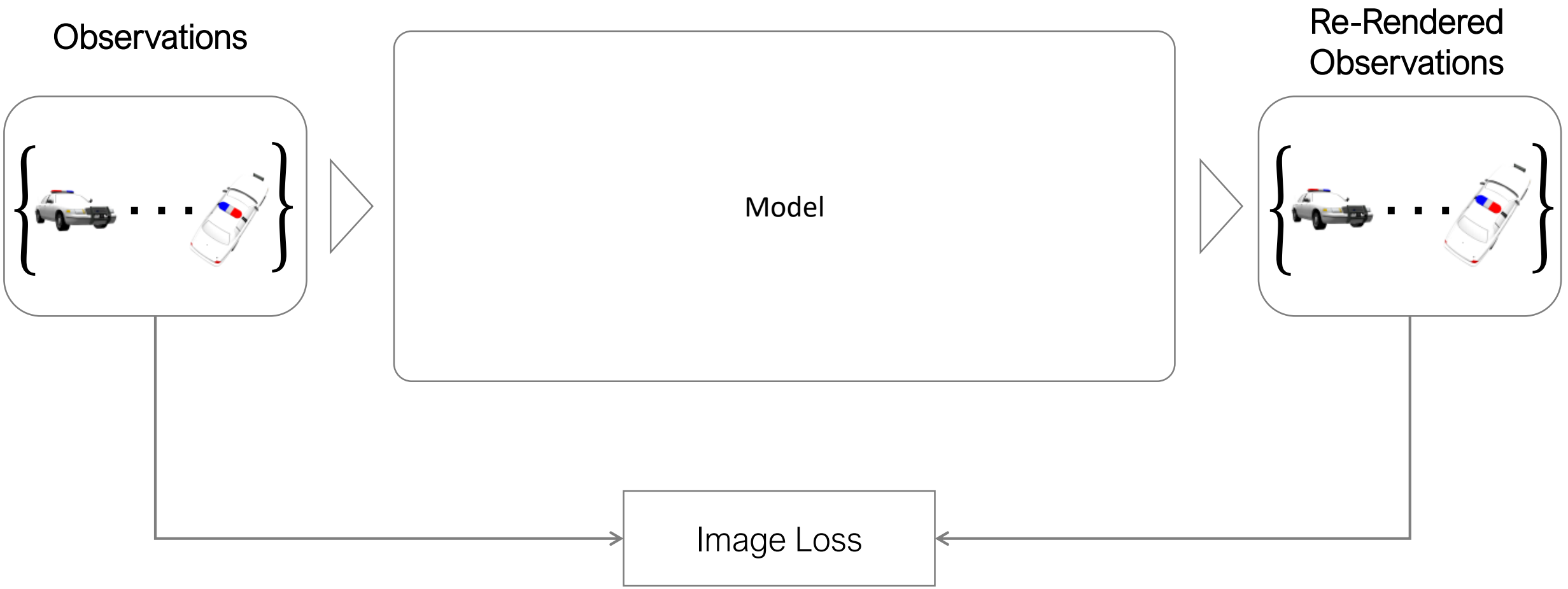
Observations
(Image, Pose, Intrinsics)



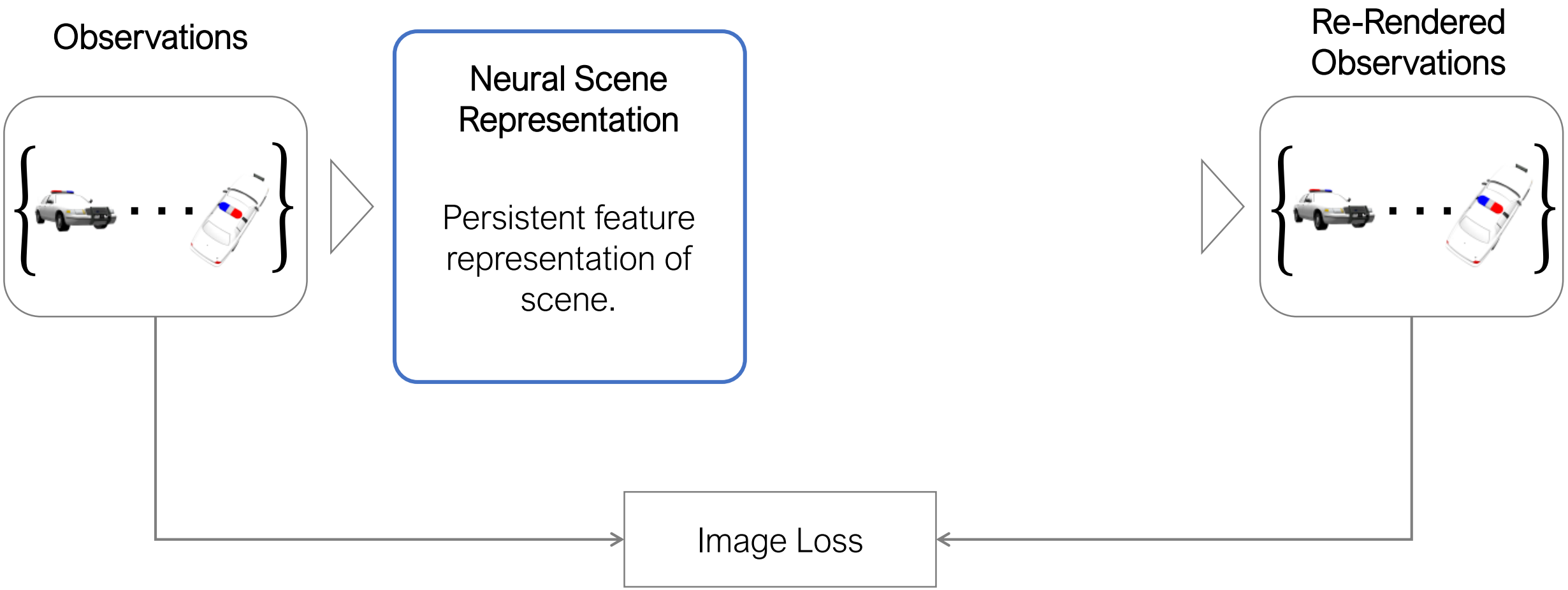
What can we learn about underlying 3D scenes?

Vision: Learn rich representations just by watching video!

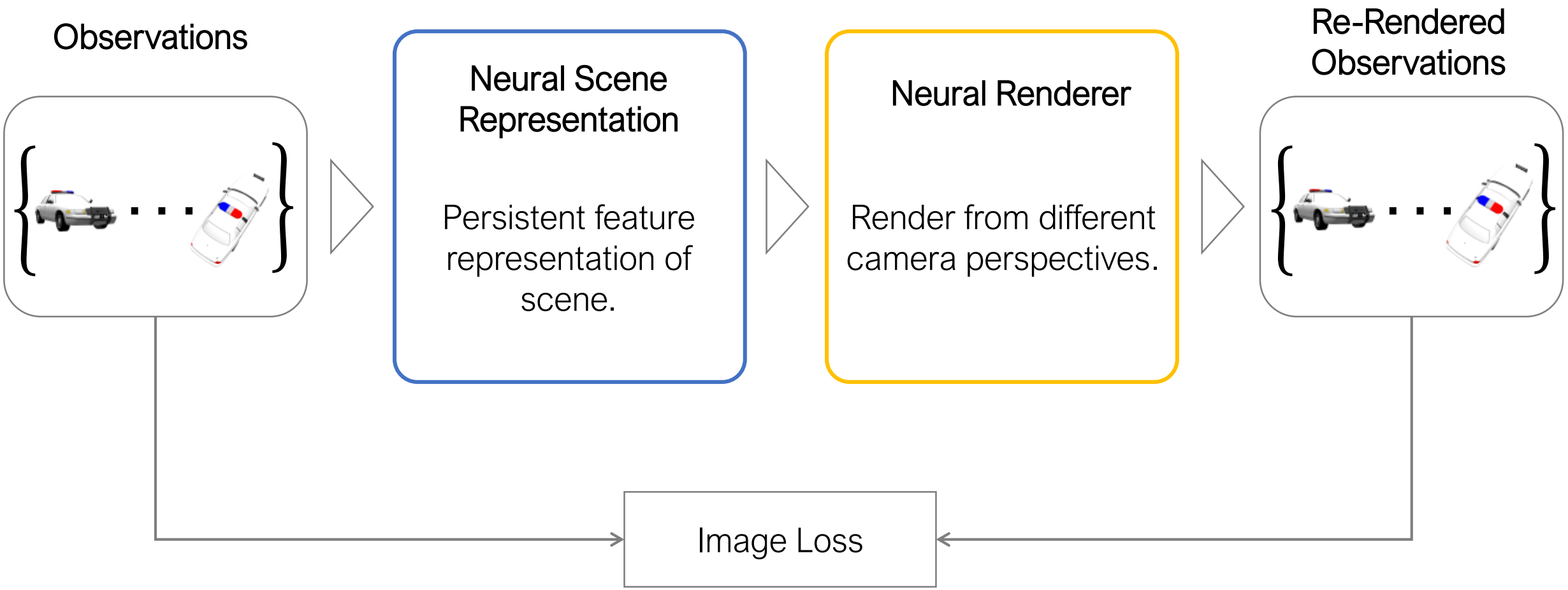
Self-supervised Scene Representation Learning



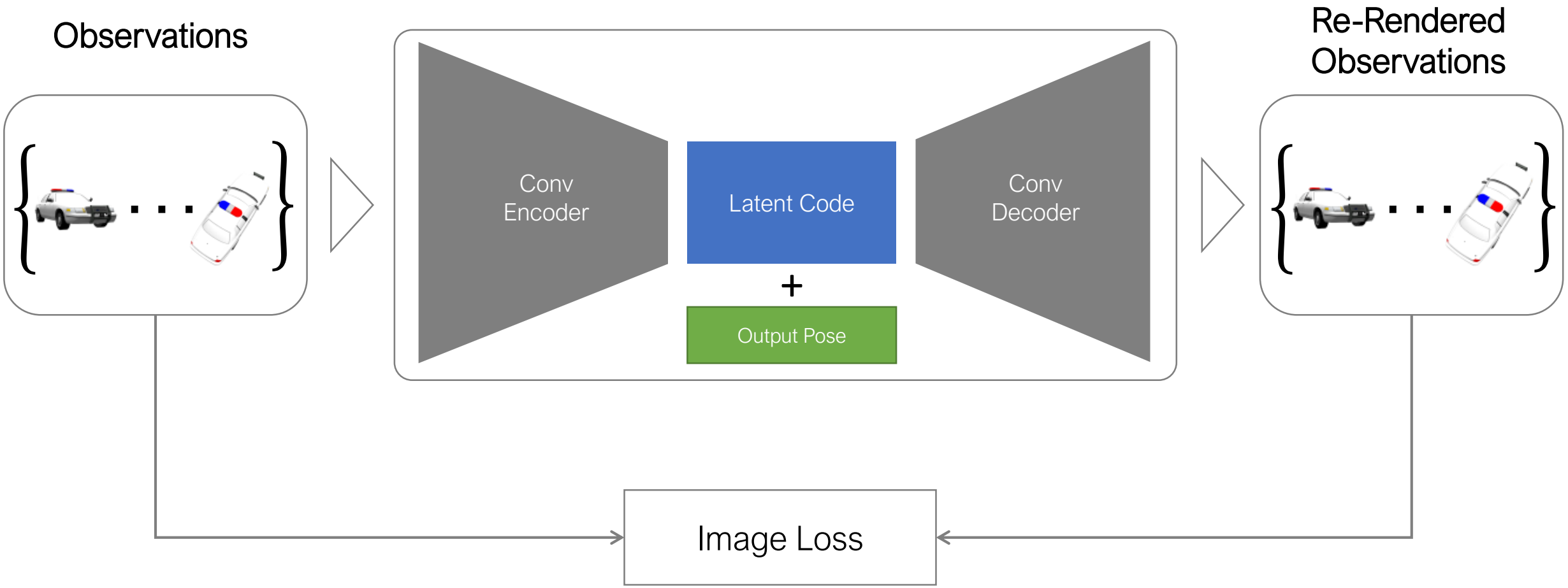
Self-supervised Scene Representation Learning



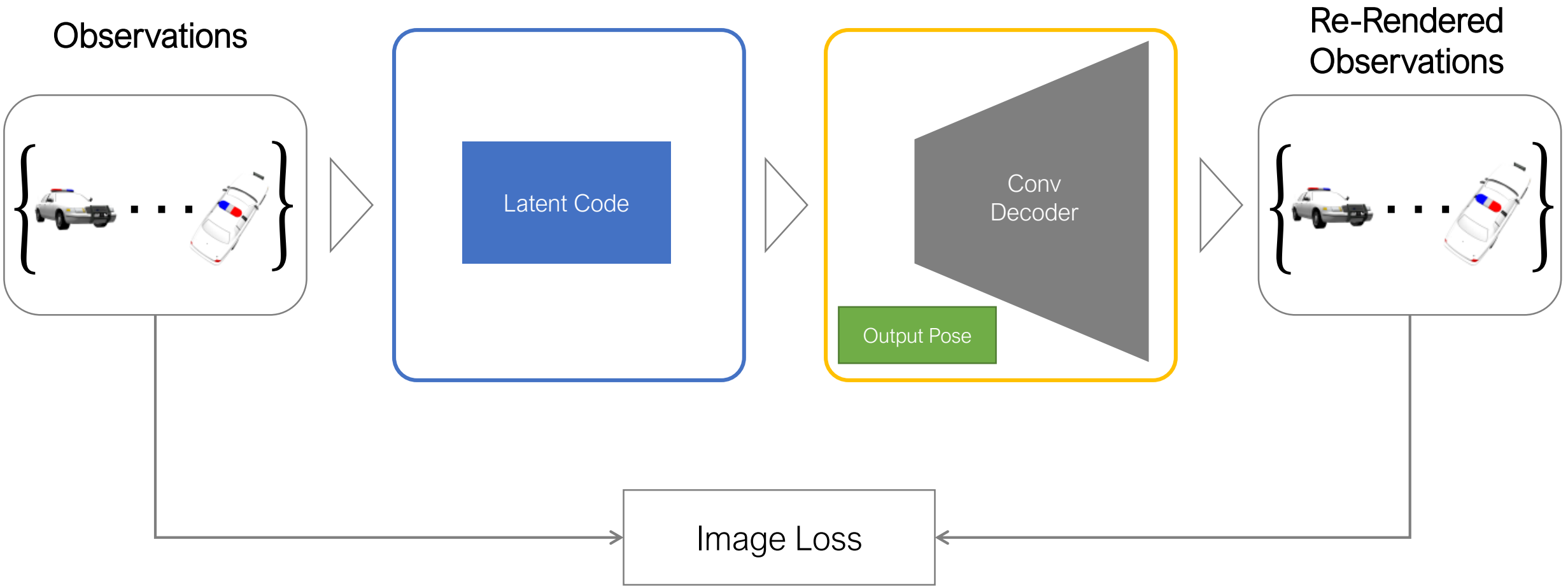
Self-supervised Scene Representation Learning



2D baseline: Autoencoder



2D baseline: Autoencoder

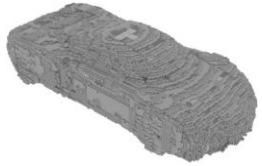
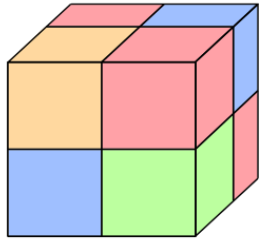


Doesn’t capture 3D properties of scenes.

Trained on ~2500 shapenet cars with 50 observations each.

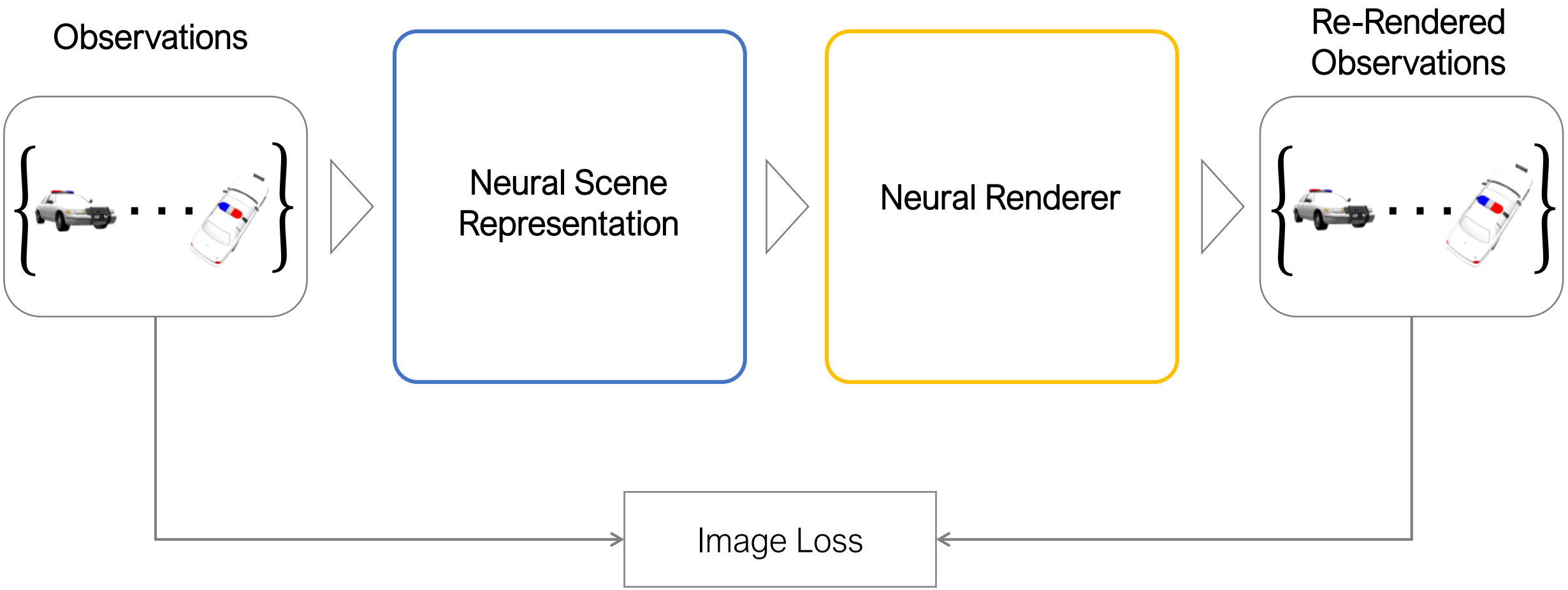
Need 3D inductive bias!

Related Work

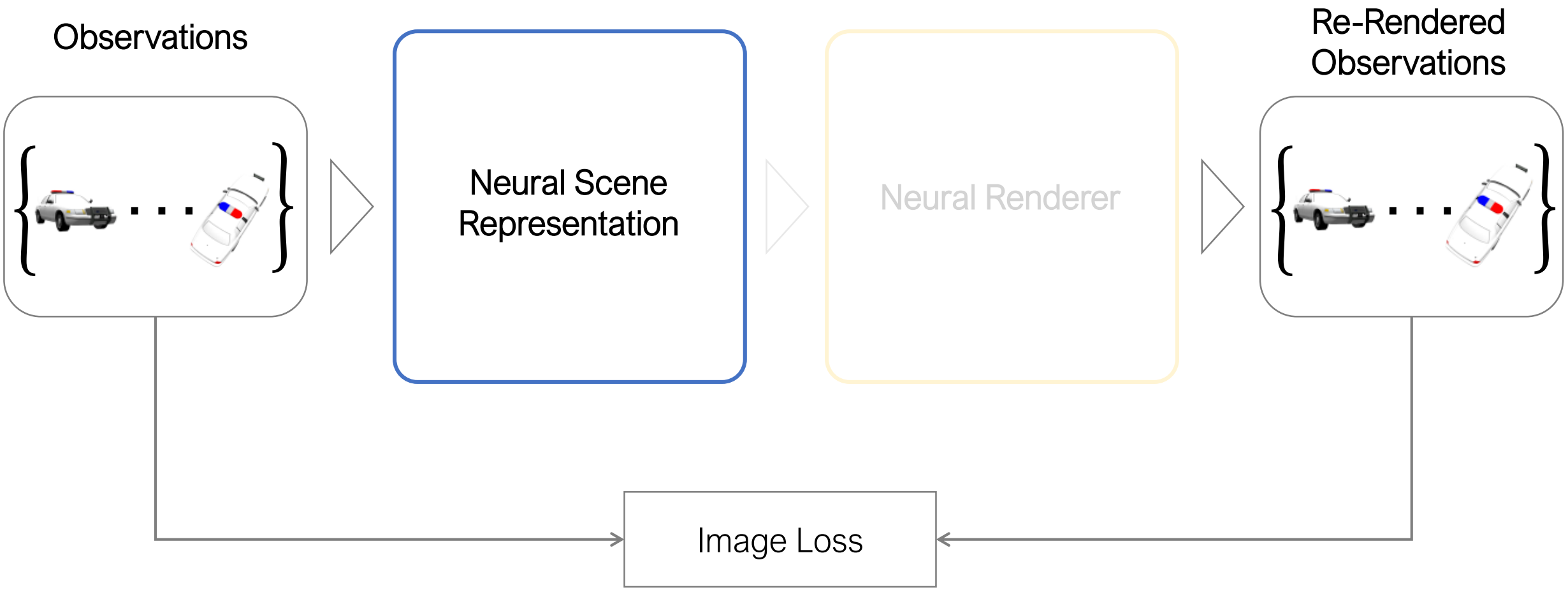
	3D inductive bias / 3D structure	Self-supervised with posed images
 **Scene Representation Learning** Tatarchenko et al., 2015 Worrall et al., 2017 Eslami et al., 2018 …	✗	✓
 **2D Generative Models** Goodfellow et al., 2014 Kingma et al., 2013 Kingma et al., 2018 …	✗	✓
 **3D Computer Vision** Choy et al., 2016 Huang et al., 2018 Park et al., 2018 …	✓	✗
 **Voxel-based Representations** Sitzmann et al., 2019 Lombardi et al., 2019 Phuoc et al., 2019 …	✓	✓

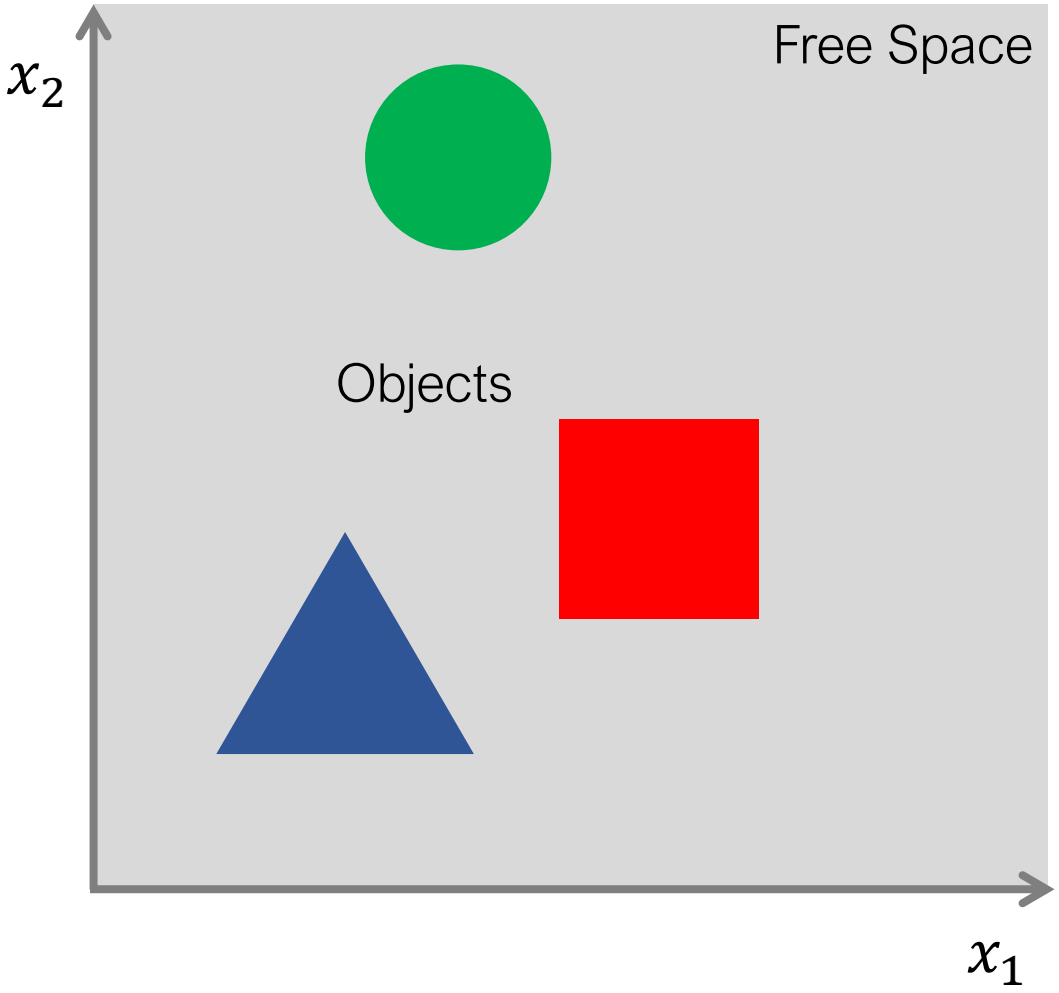
- Memory inefficient: $O(n^3)$.
- Doesn't parameterize scene surfaces smoothly.
- Generalization is hard.

Scene Representation Networks

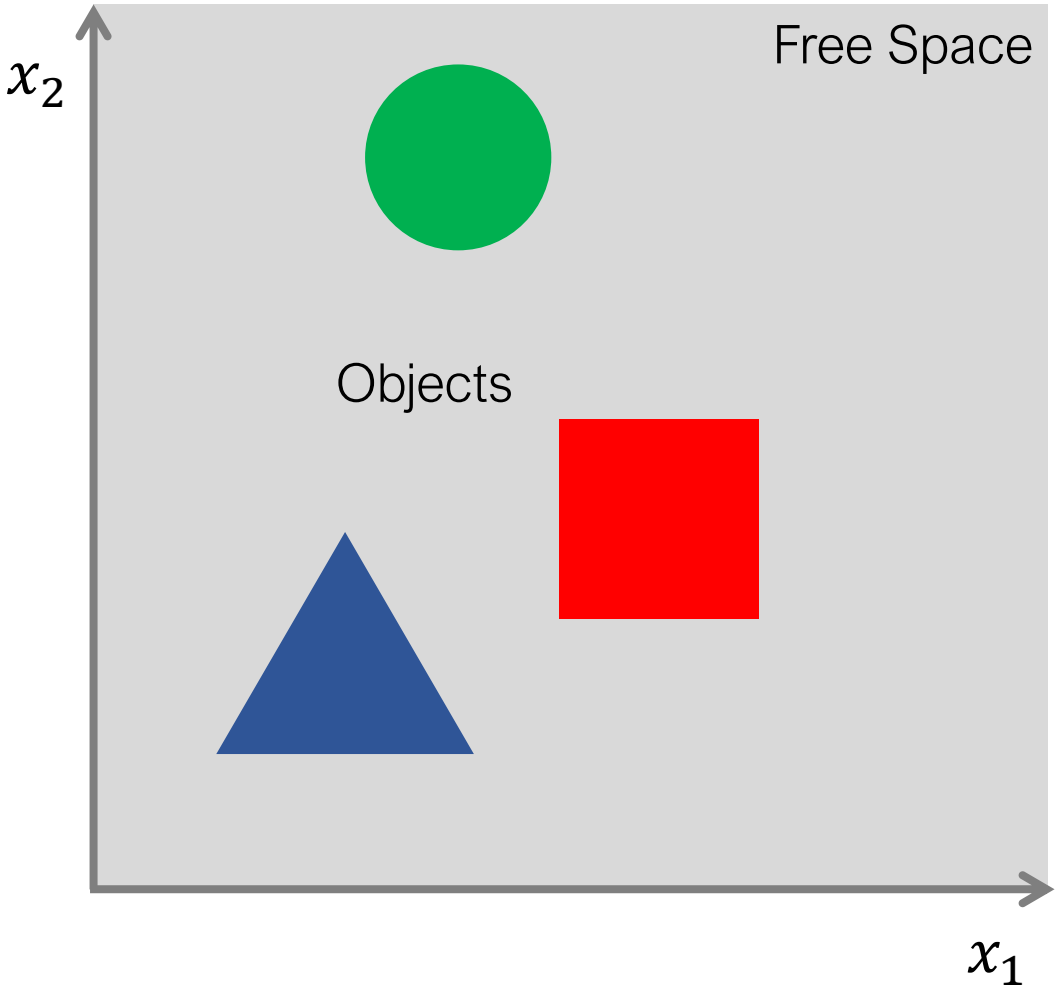


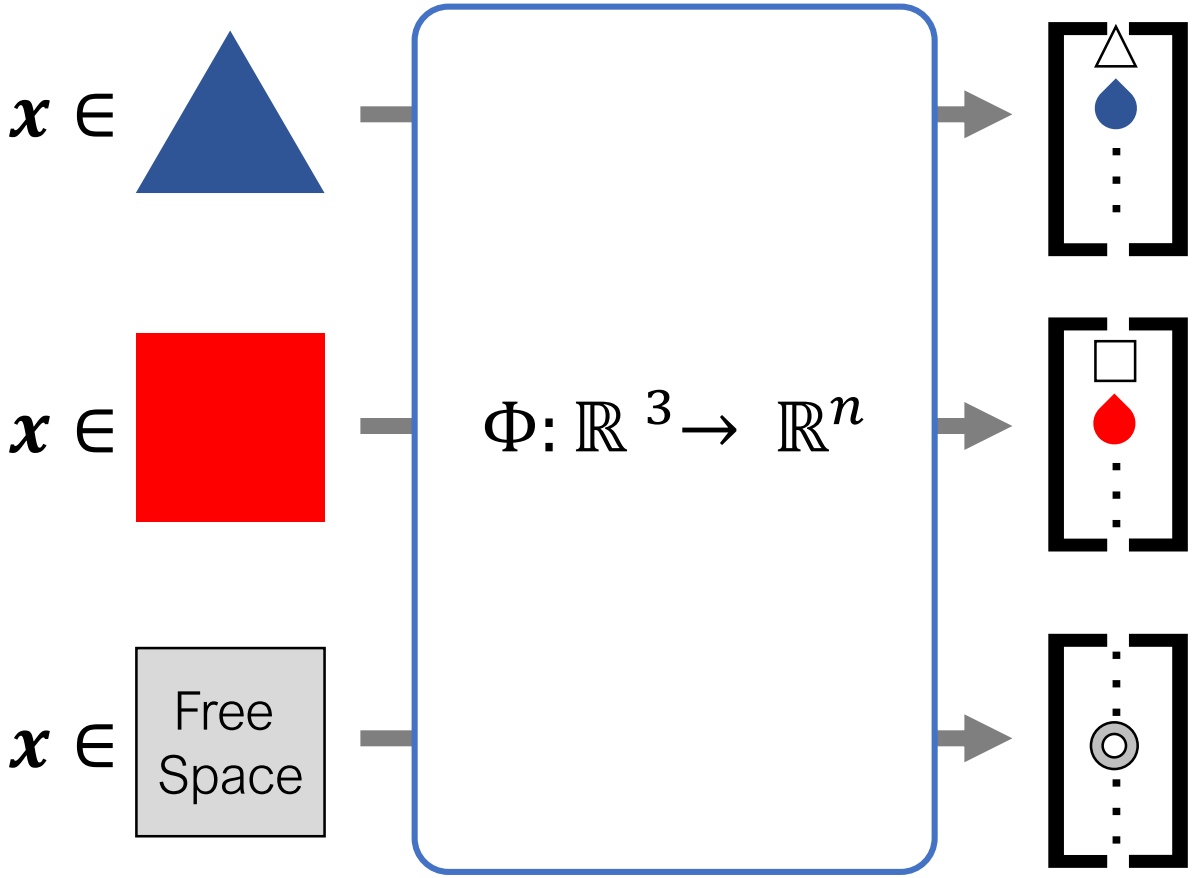
Scene Representation Networks



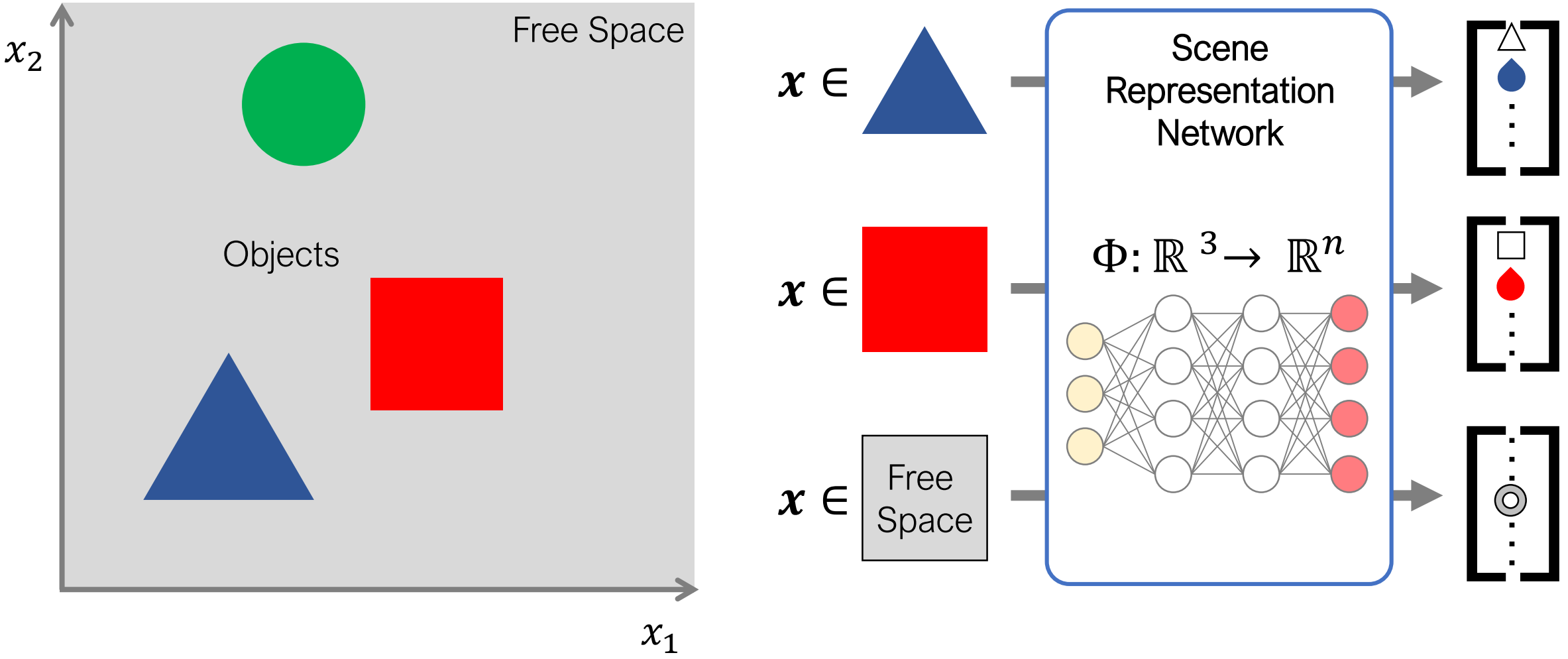

Free Space
x_2
Objects
x_1

Model scene as function Φ that maps coordinates to features.





Scene Representation Network parameterizes Φ as MLP.

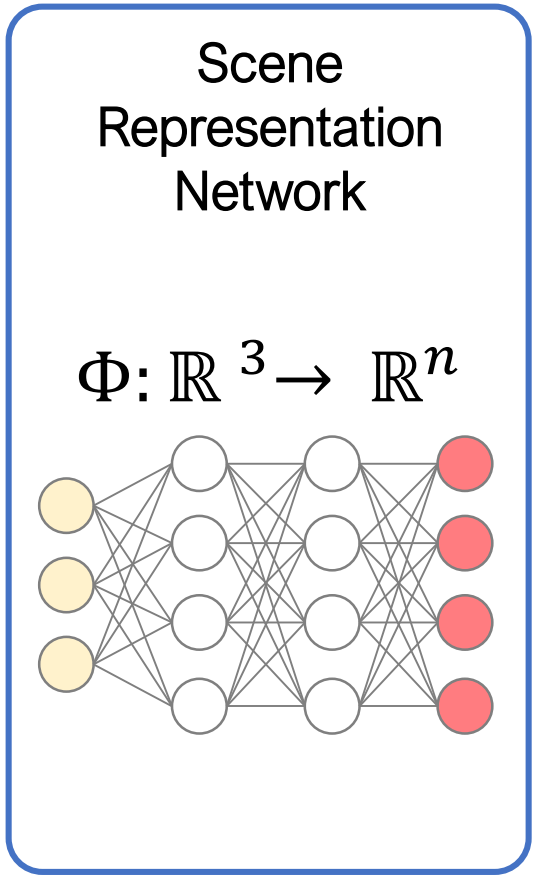


Scene Representation Network parameterizes Φ as MLP.

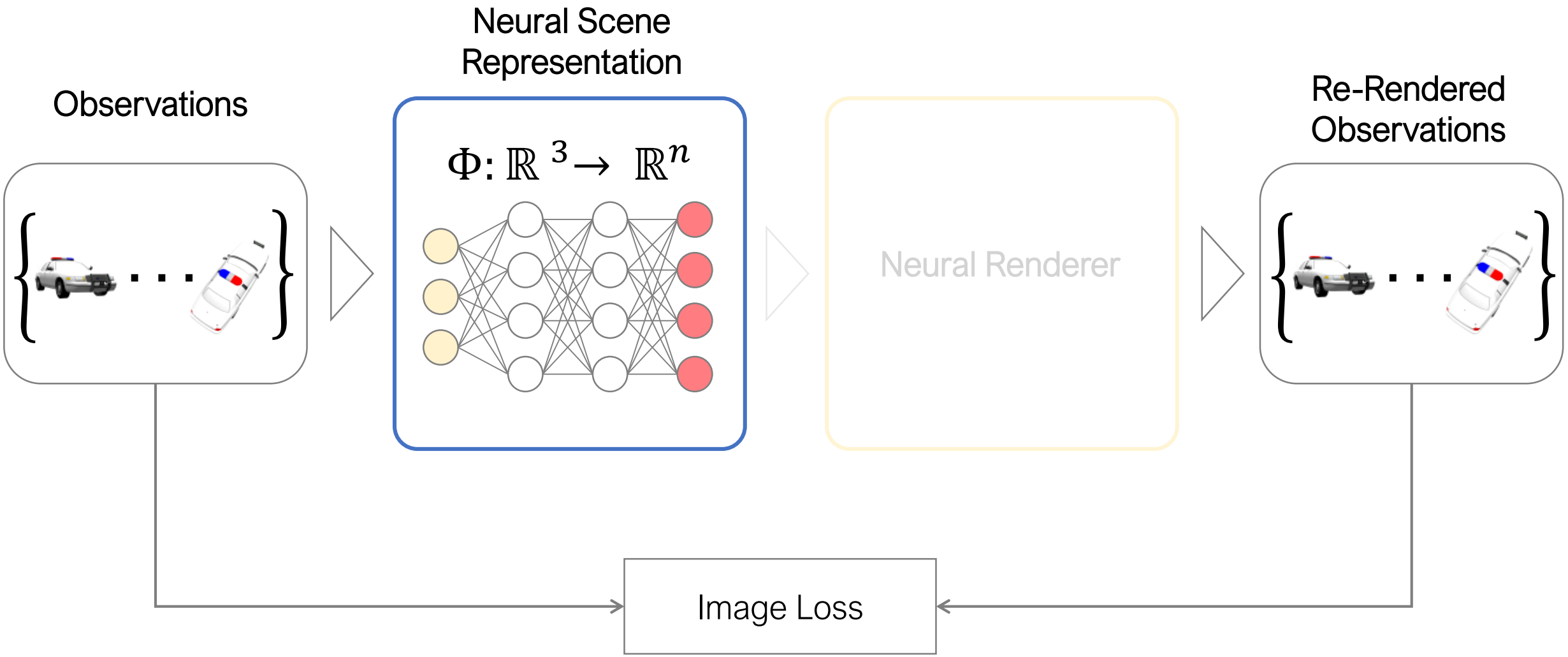
Can sample anywhere,
at arbitrary resolutions.

Parameterizes scene
surfaces smoothly.

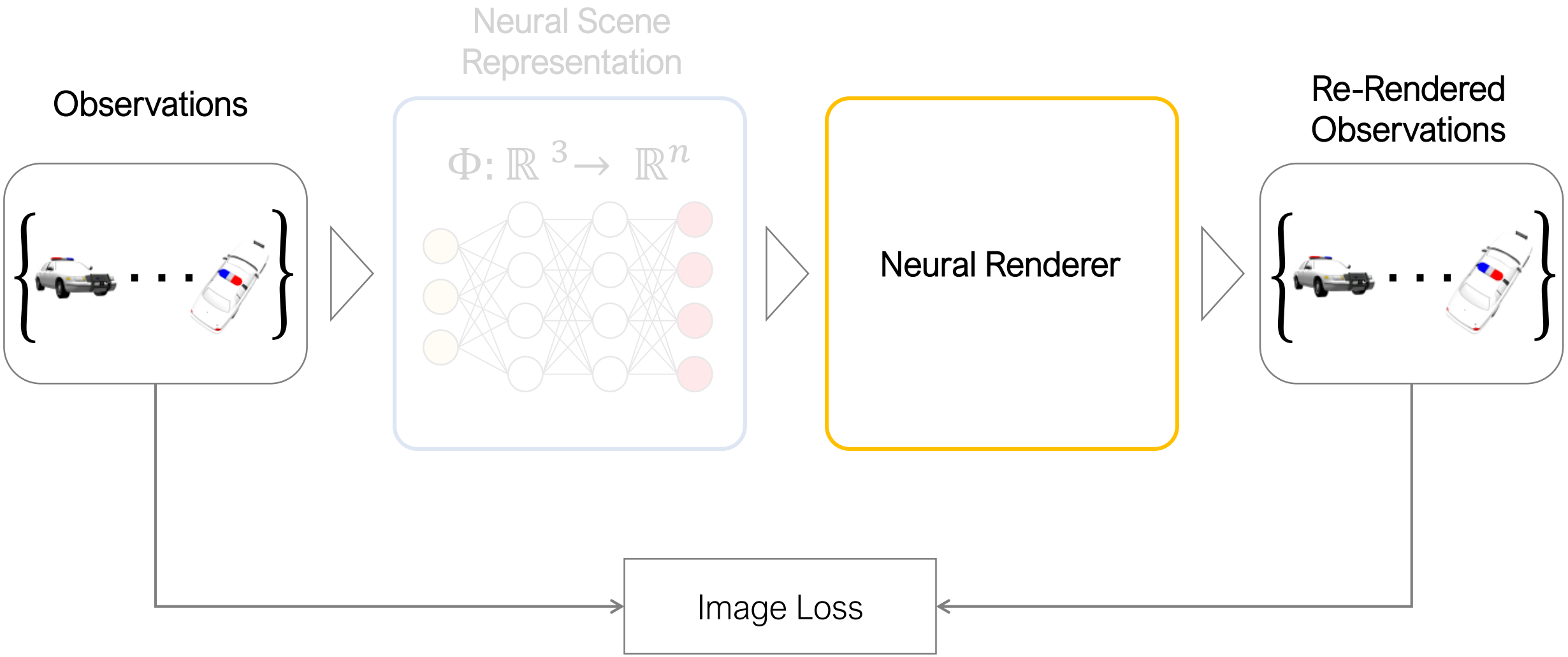
Memory scales with scene
complexity.



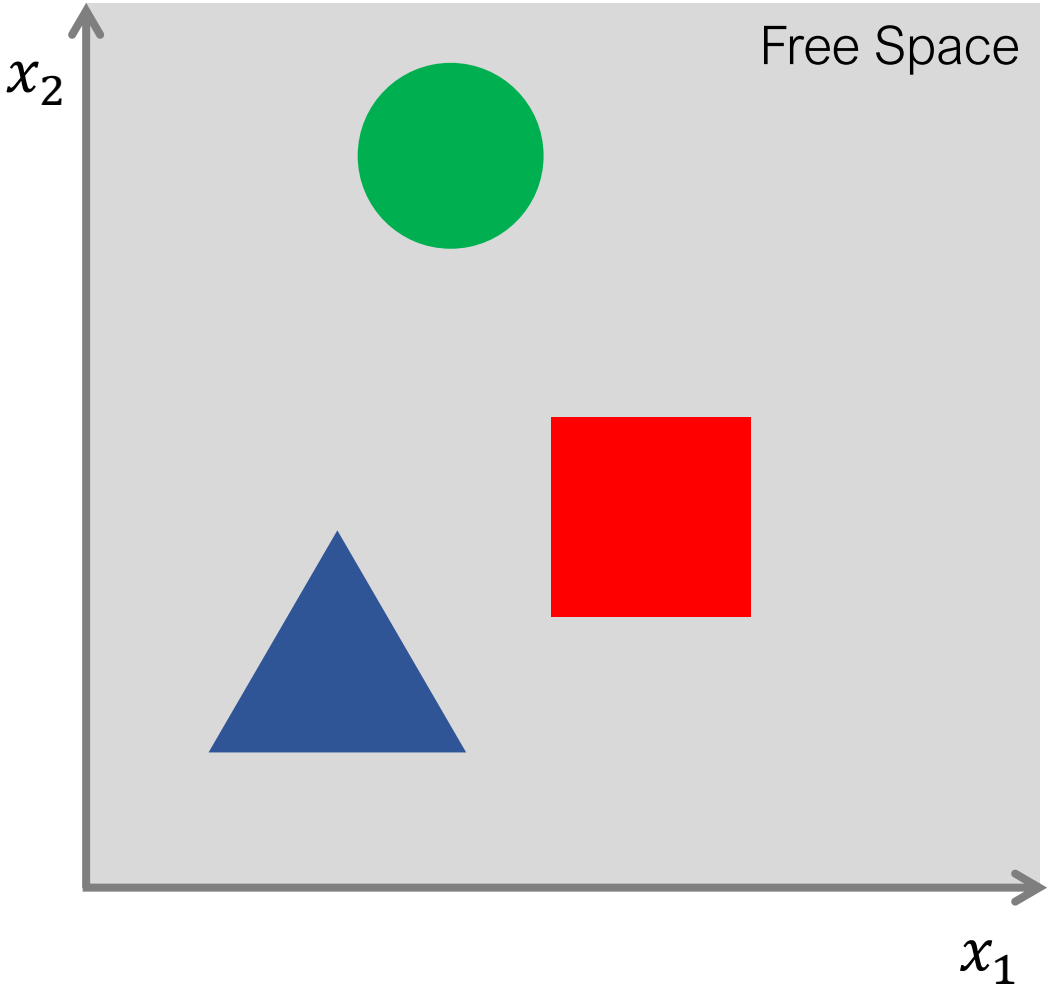
Scene Representation Networks



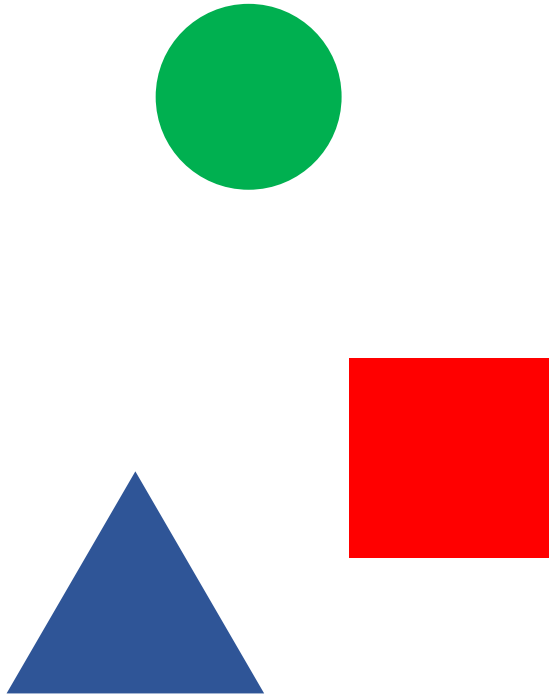
Scene Representation Networks



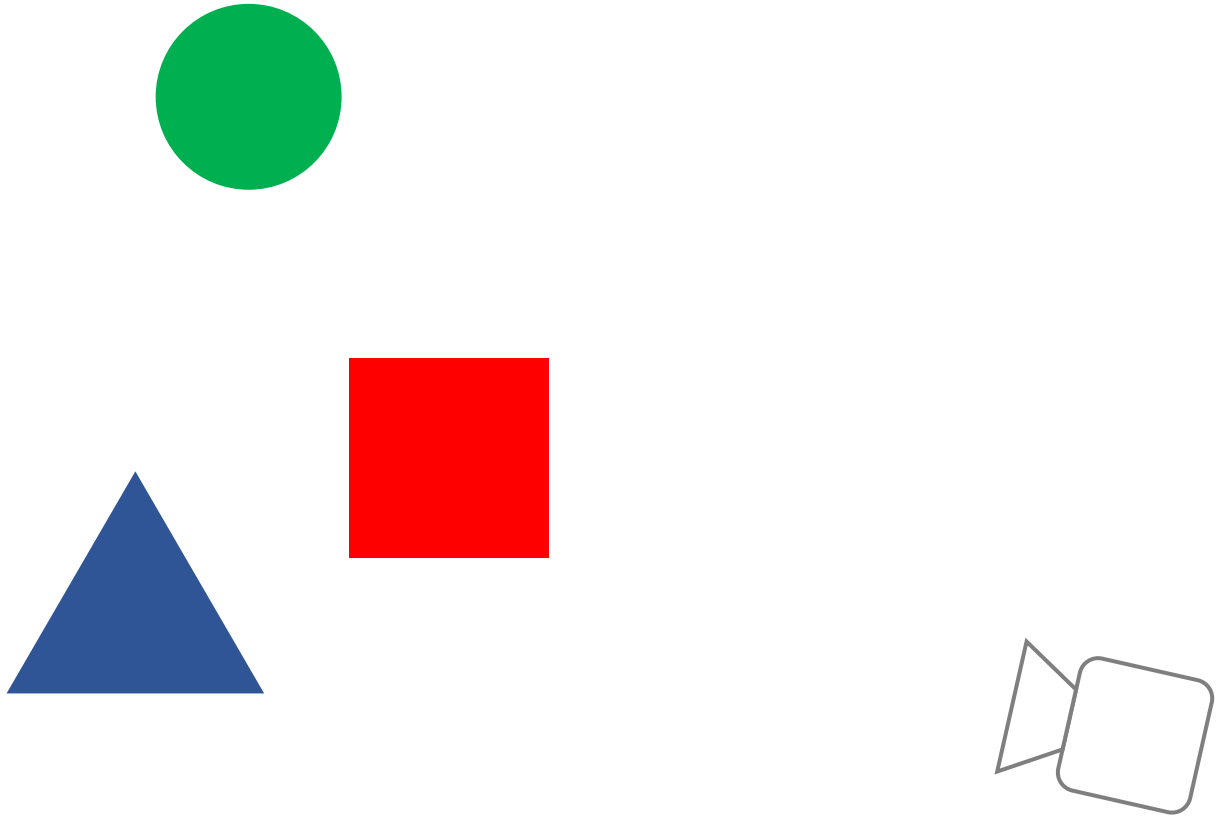
Neural Renderer.



Neural Renderer.

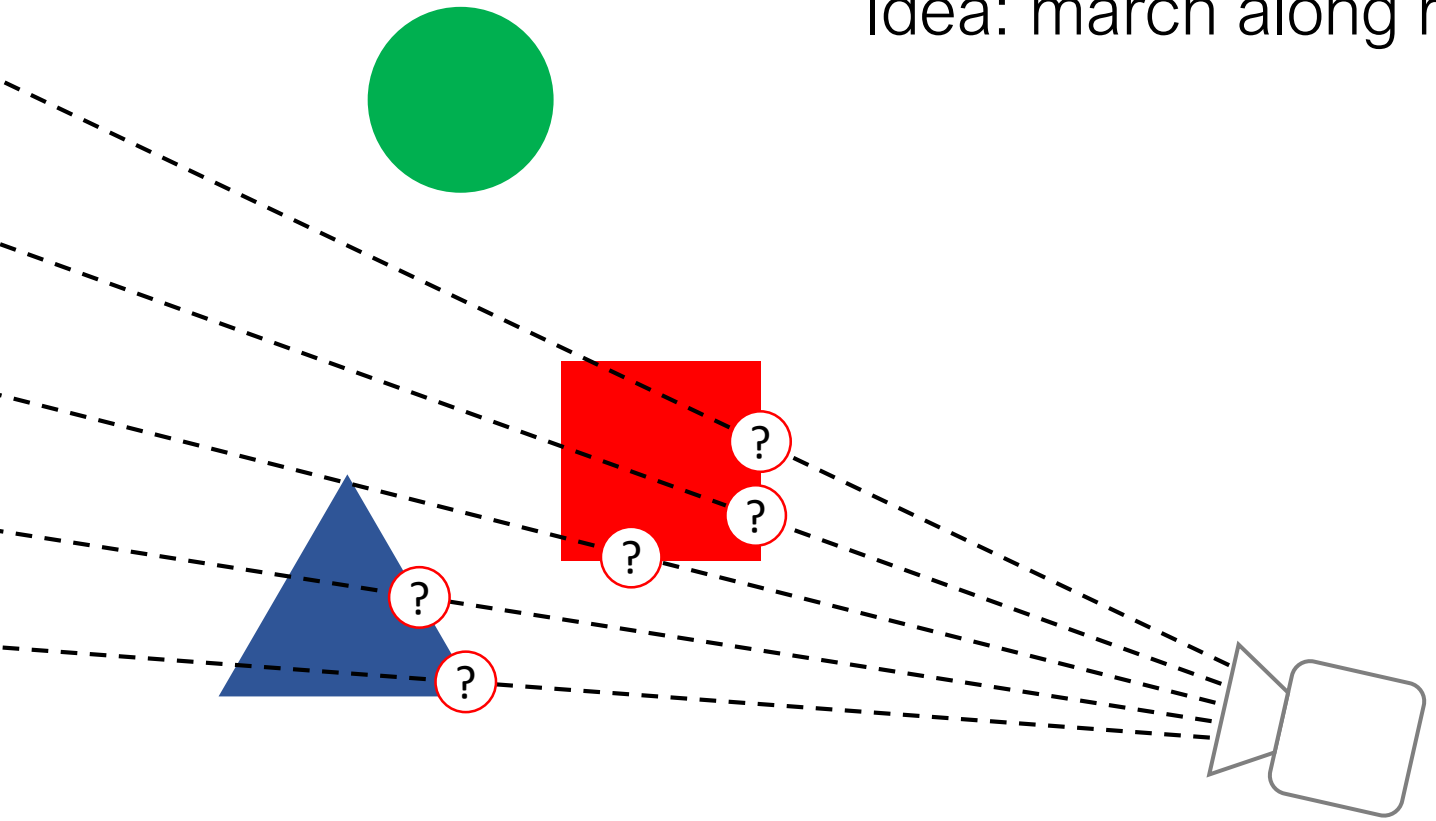


Neural Renderer.

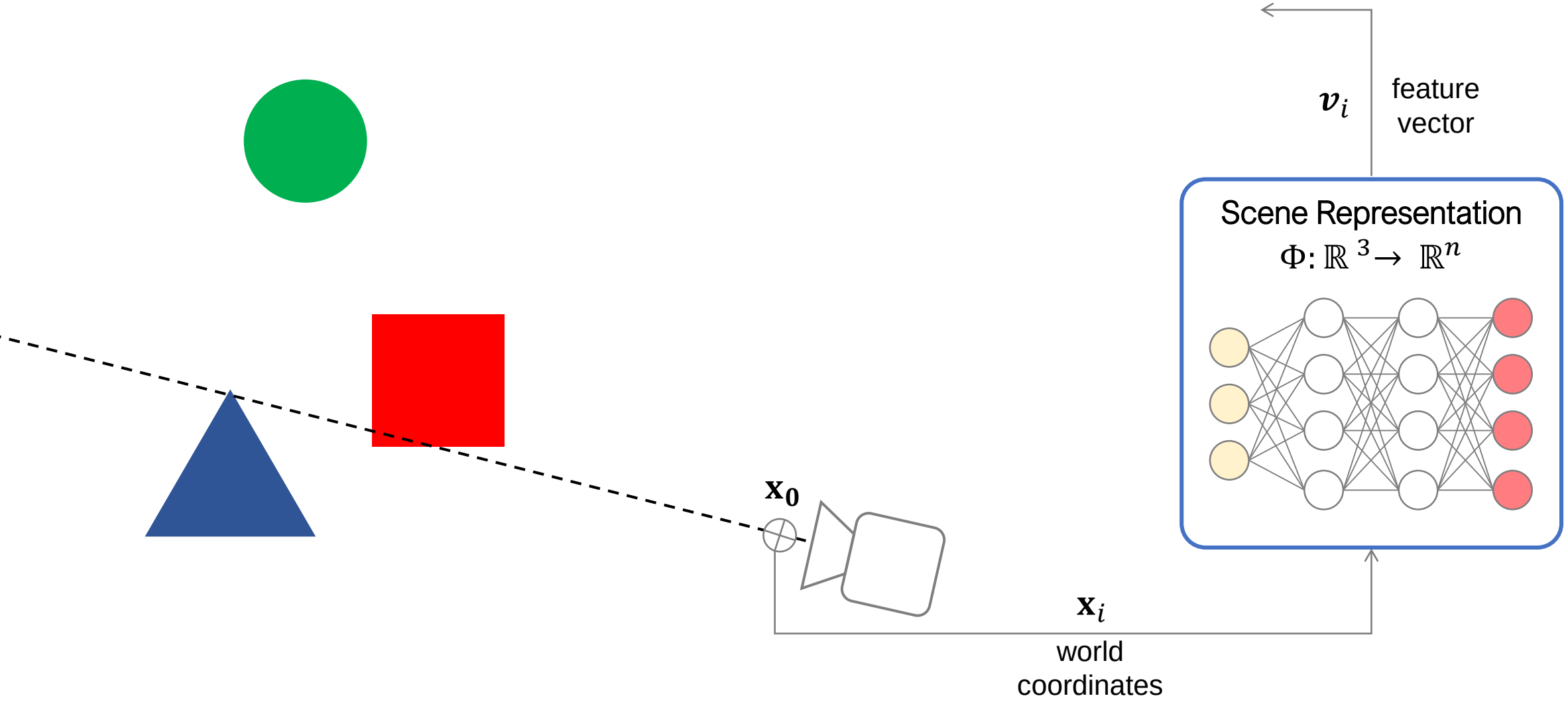


Neural Renderer Step 1: Intersection Testing.

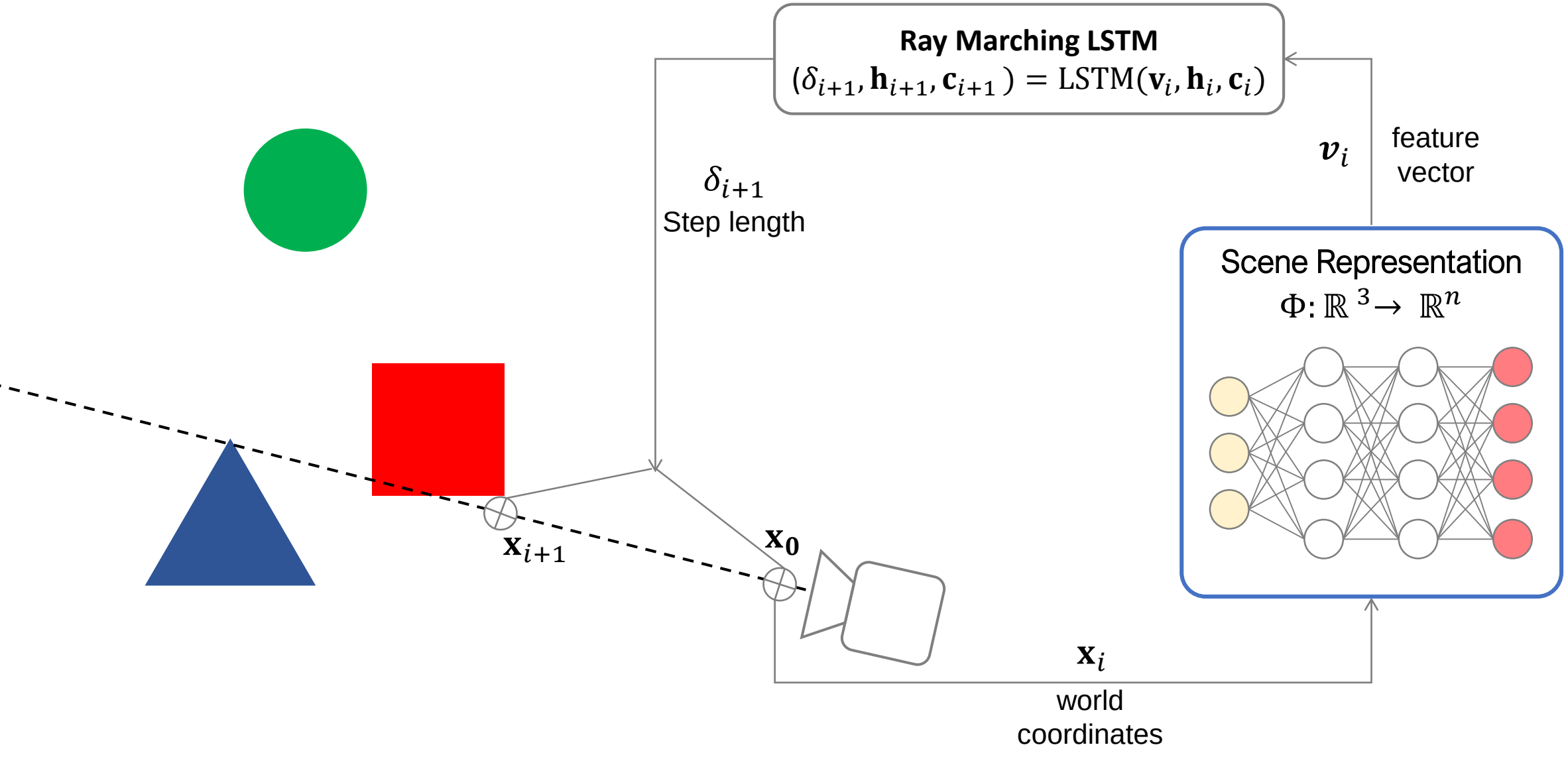
Idea: march along ray until arrived at surface.



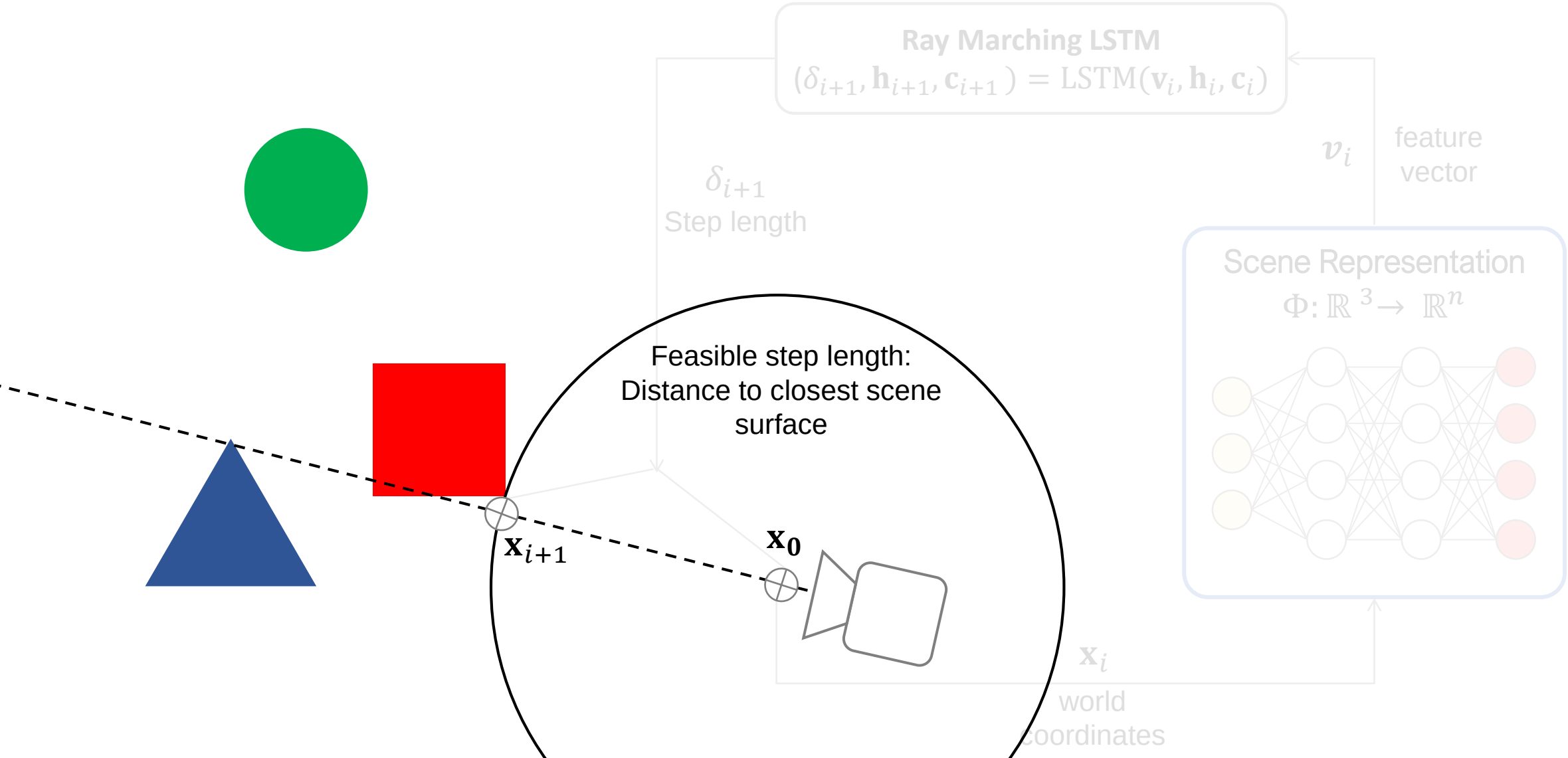
Neural Renderer Step 1: Intersection Testing.



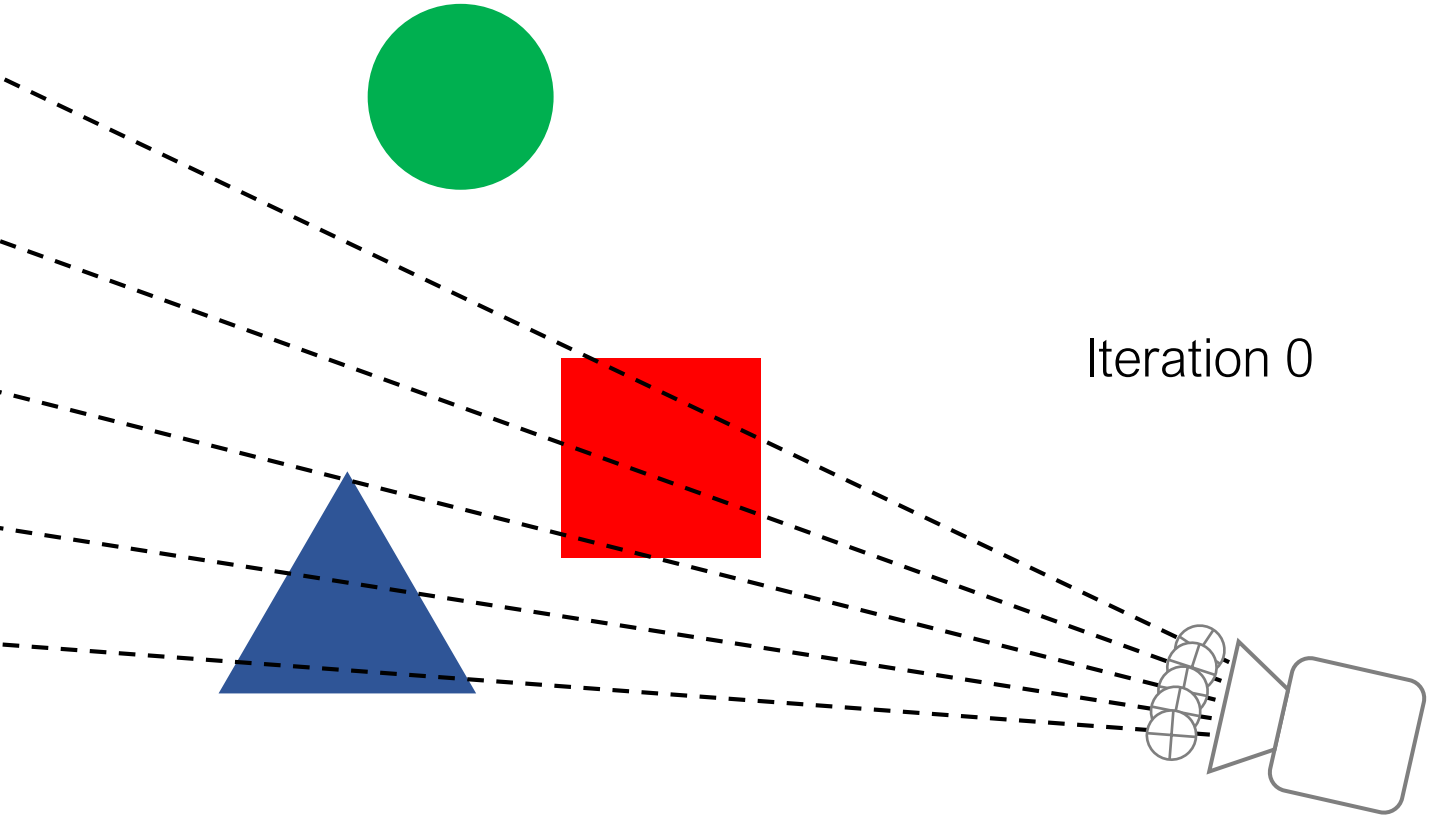
Neural Renderer Step 1: Intersection Testing.



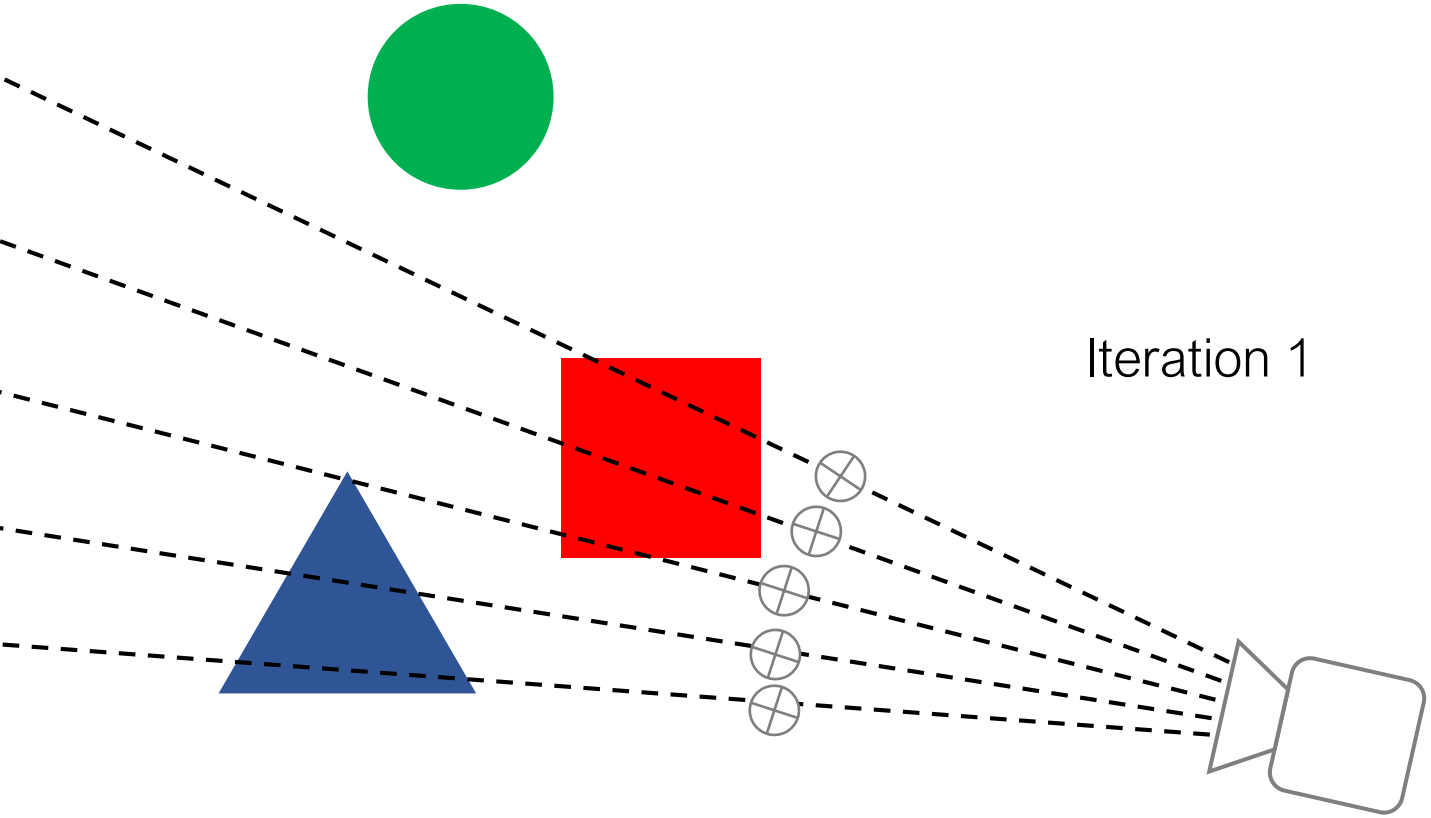
Neural Renderer Step 1: Intersection Testing.



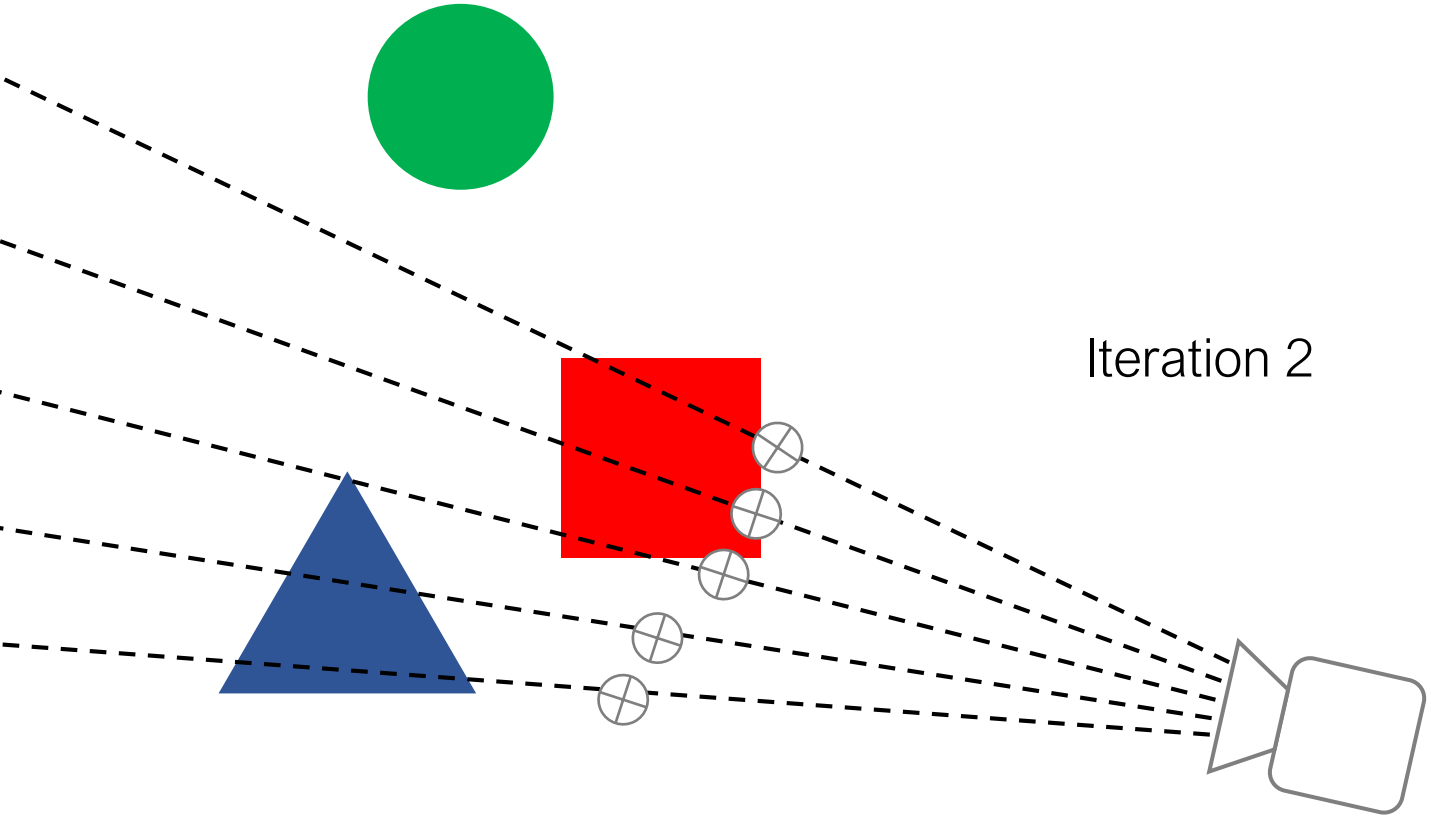
Neural Renderer Step 1: Intersection Testing.



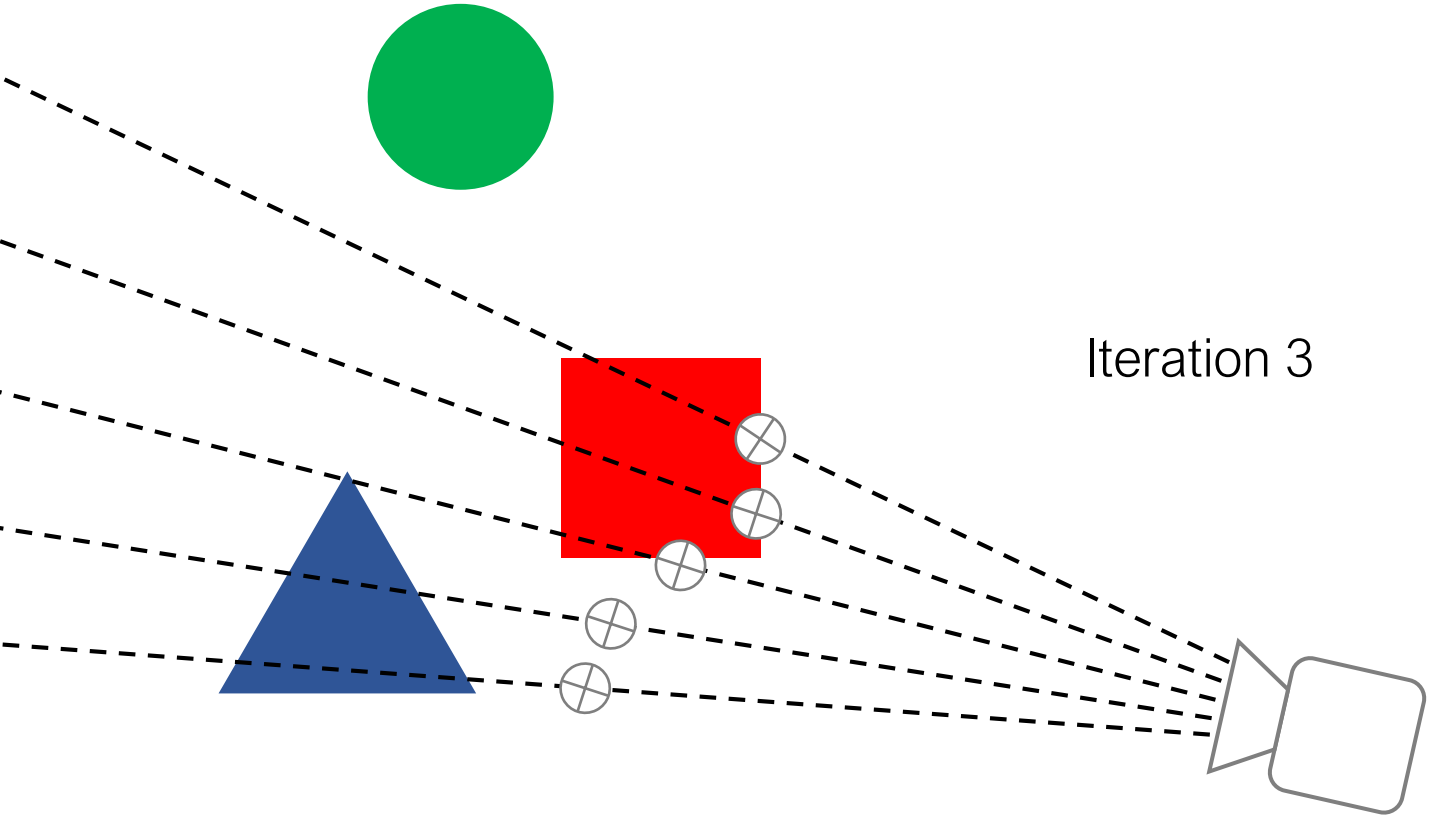
Neural Renderer Step 1: Intersection Testing.



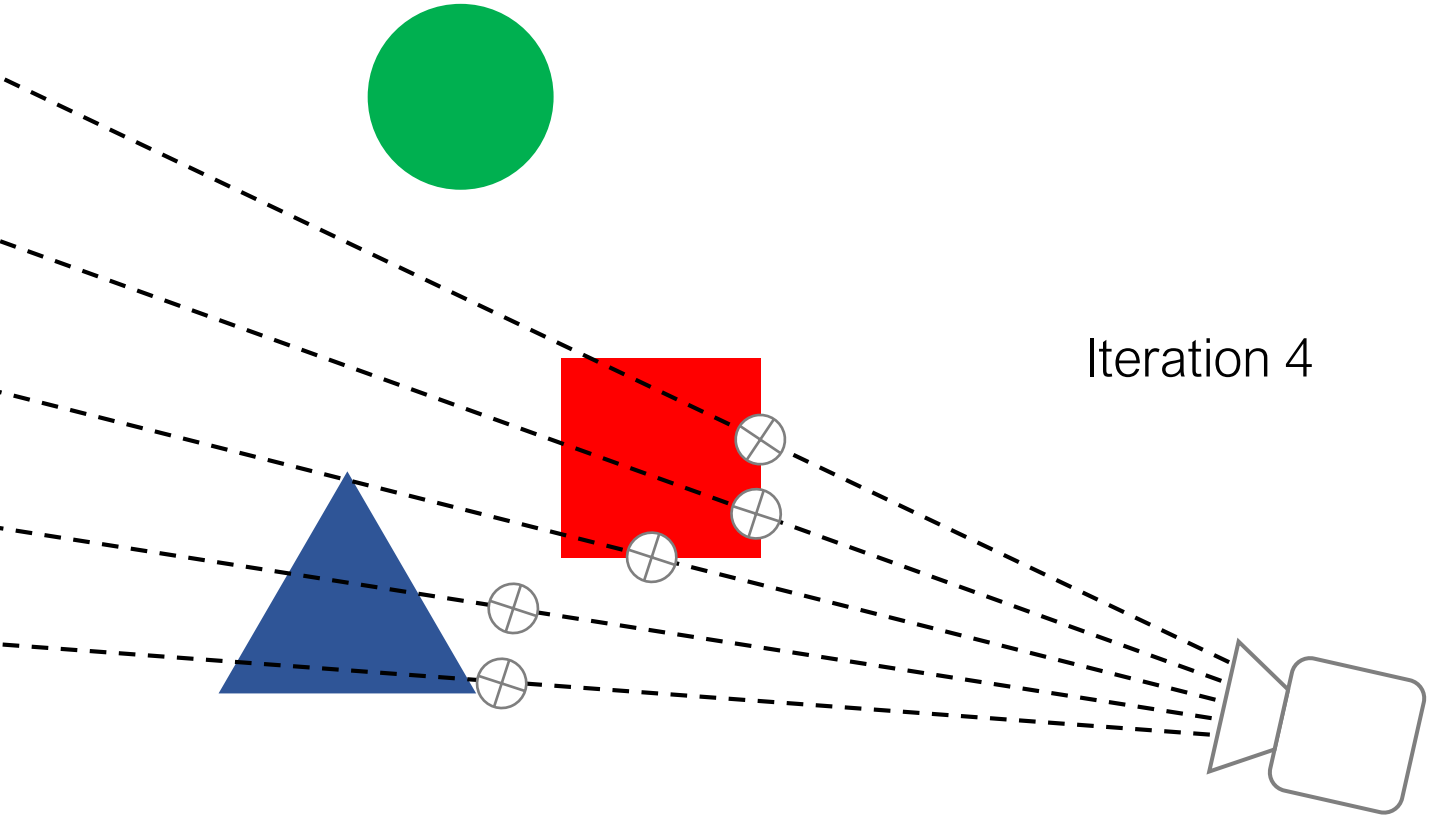
Neural Renderer Step 1: Intersection Testing.



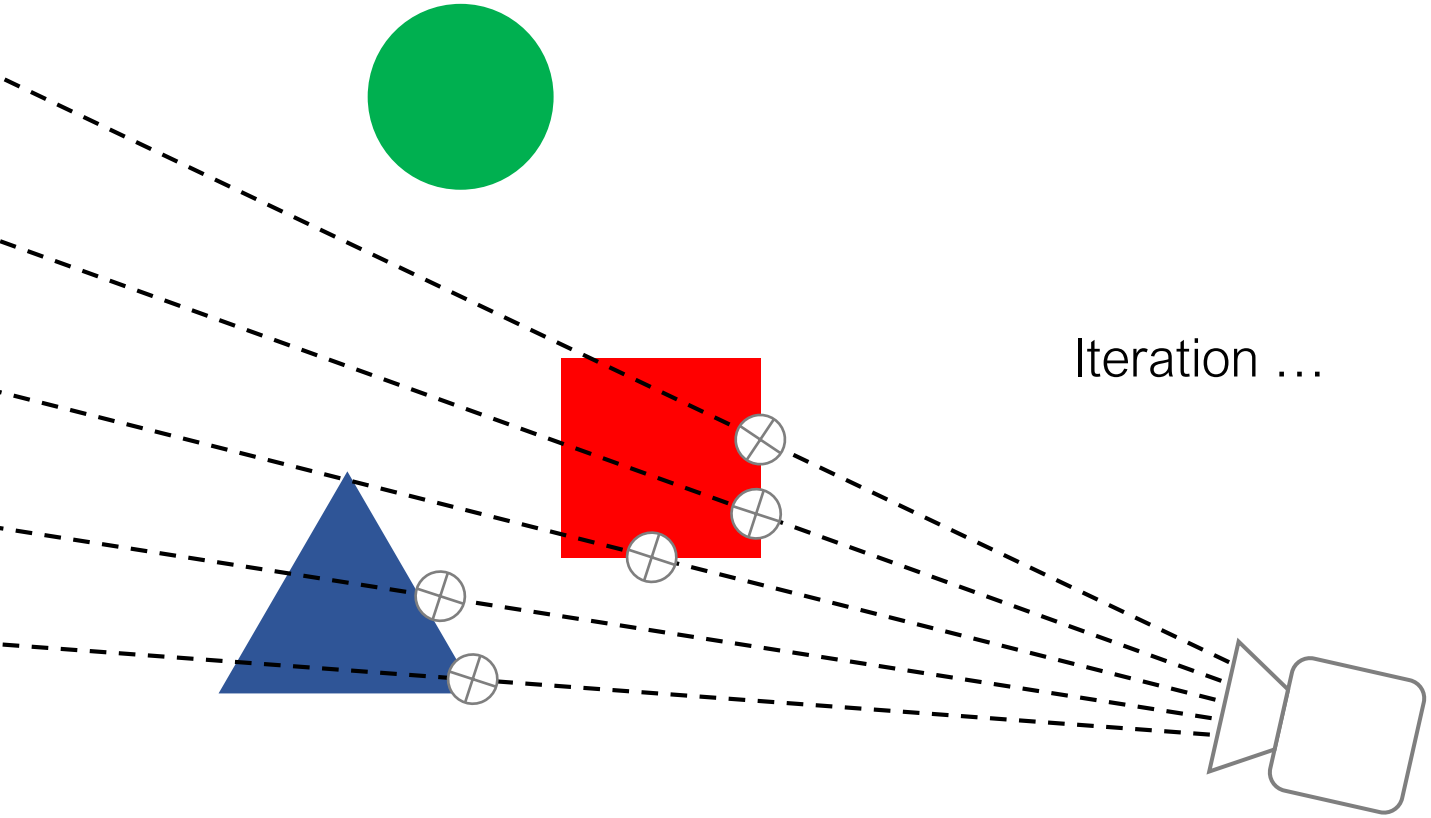
Neural Renderer Step 1: Intersection Testing.



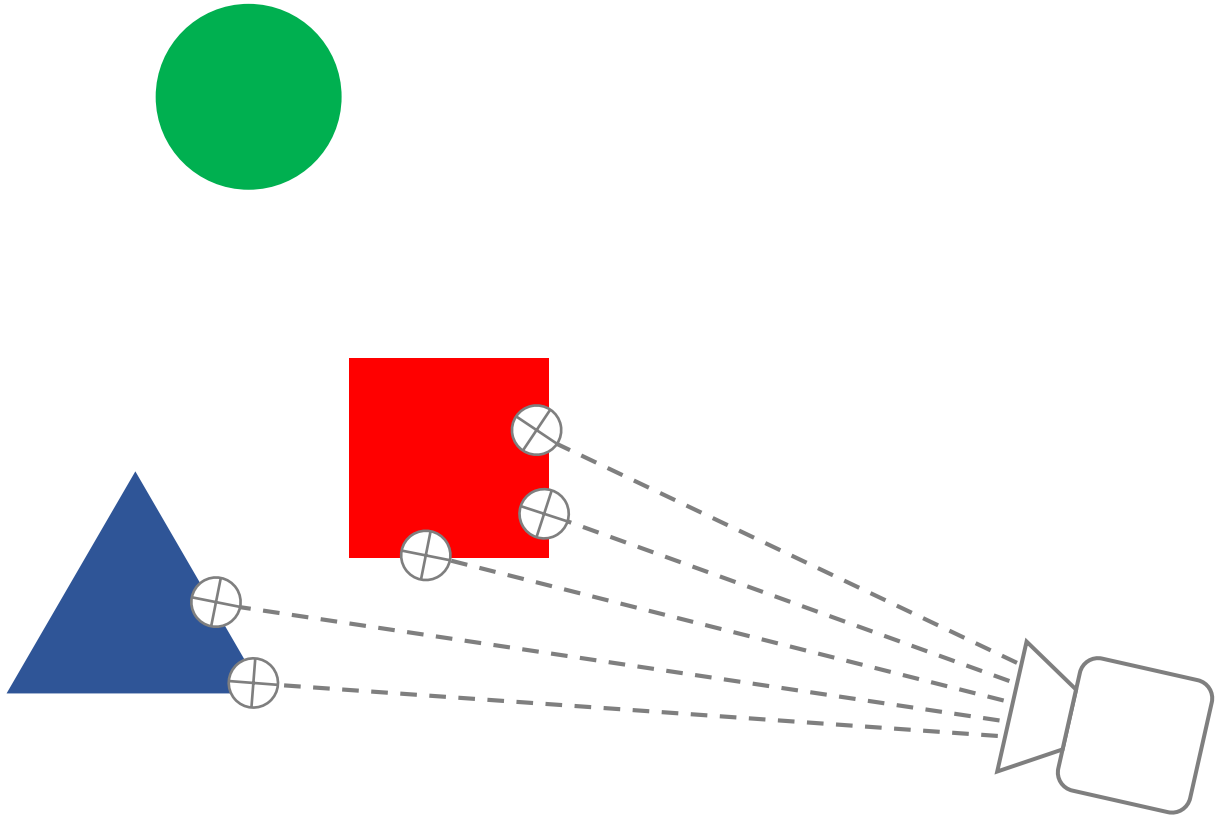
Neural Renderer Step 2: Color Generation



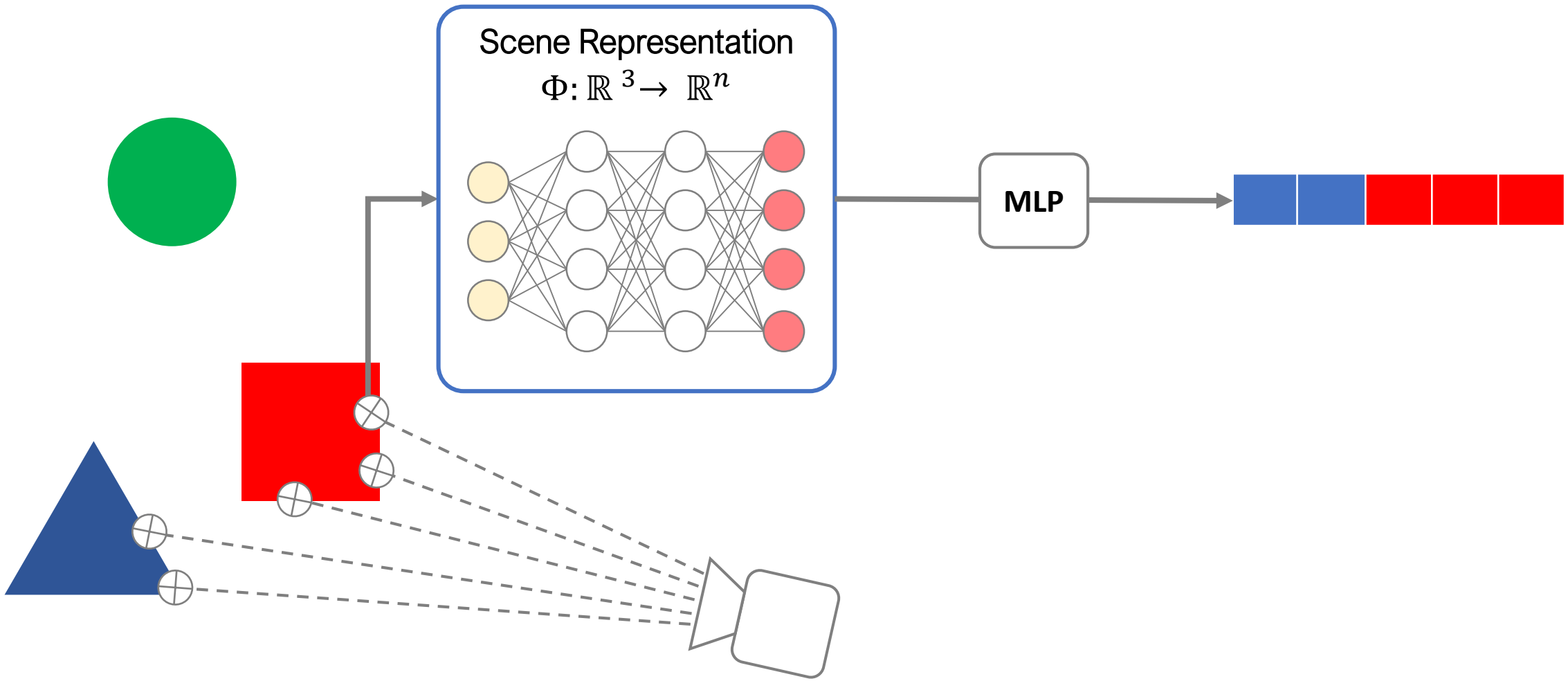
Neural Renderer Step 1: Intersection Testing.



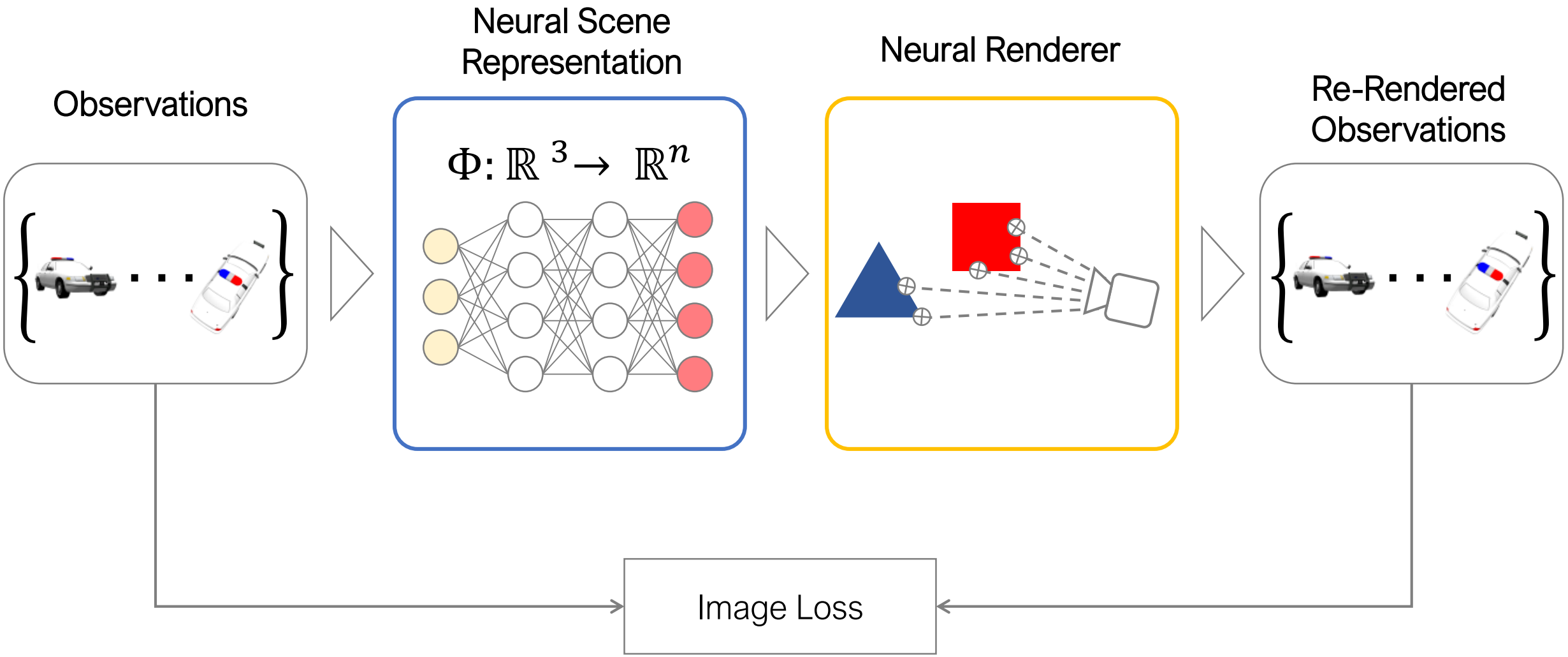
Neural Renderer Step 1: Intersection Testing.



Neural Renderer Step 2: Color Generation



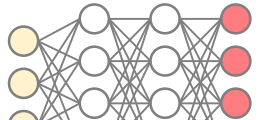
Can now train end-to-end with posed images only!



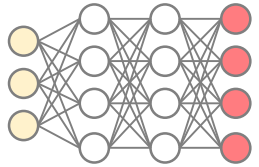
Generalizing across a class of scenes

Each scene represented by its own SRN.

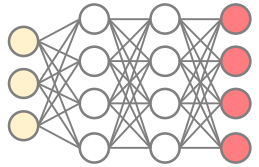
parameters $\phi_0 \in \mathbb{R}^l$



parameters $\phi_1 \in \mathbb{R}^l$

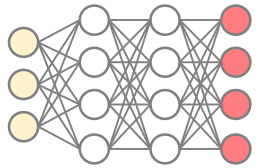


parameters $\phi_2 \in \mathbb{R}^l$



○ ○ ○

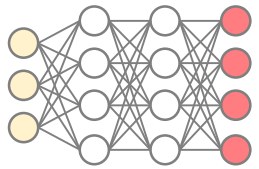
parameters $\phi_n \in \mathbb{R}^l$



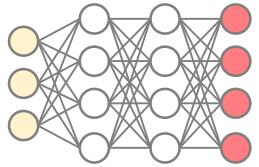
Each scene represented by its own SRN.

ϕ_i live on k-dimensional subspace of $\mathbb{R}^l$, $k < l$.

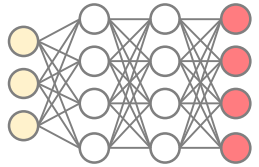
parameters $\phi_0 \in \mathbb{R}^l$



parameters $\phi_1 \in \mathbb{R}^l$

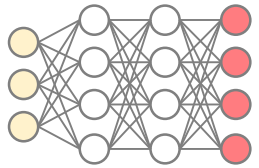


parameters $\phi_2 \in \mathbb{R}^l$



○ ○ ○

parameters $\phi_n \in \mathbb{R}^l$



Each scene represented by its own SRN.

embedding $z_0 \in \mathbb{R}^k$

embedding $z_1 \in \mathbb{R}^k$

embedding $z_2 \in \mathbb{R}^k$

embedding $z_n \in \mathbb{R}^k$

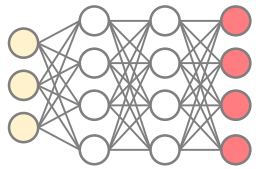
Represent each scene with low-dimensional embedding

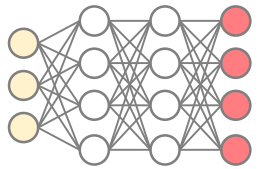
parameters $\phi_0 \in \mathbb{R}^l$

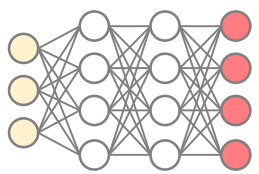
parameters $\phi_1 \in \mathbb{R}^l$

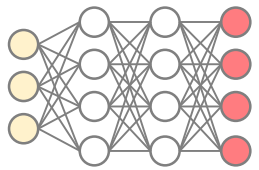
parameters $\phi_2 \in \mathbb{R}^l$

parameters $\phi_n \in \mathbb{R}^l$









Each scene represented by its own SRN.

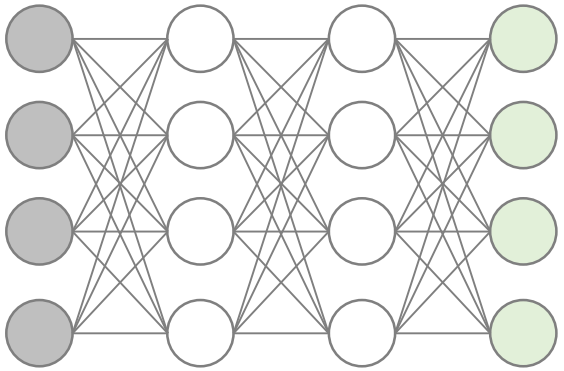
embedding $z_0 \in \mathbb{R}^k$

embedding $z_1 \in \mathbb{R}^k$

embedding $z_2 \in \mathbb{R}^k$

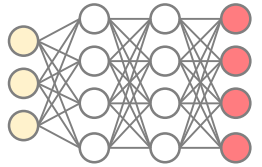
embedding $z_n \in \mathbb{R}^k$

Hypernetwork

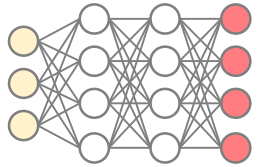


$$\Psi: \mathbb{R}^k \rightarrow \mathbb{R}^l,$$
$$z_i \mapsto \Psi(z_i) = \phi_i$$

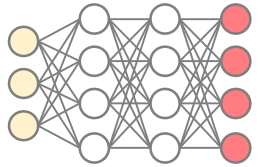
parameters $\phi_0 \in \mathbb{R}^l$



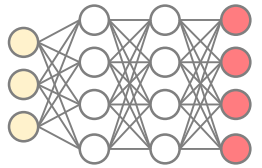
parameters $\phi_1 \in \mathbb{R}^l$



parameters $\phi_2 \in \mathbb{R}^l$



parameters $\phi_n \in \mathbb{R}^l$



Results

Novel View Synthesis – Baseline Comparison

Shapenet v2 – *single-shot reconstruction* of objects in held-out test set

Tatarchenko et al.
2015

Worrall et al.
2017

Deterministic
GQN, adapted
Eslami et al.
2018

SRNs

Training

- Shapenet cars / chairs.
- 50 observations per object.

Testing

- Cars / chairs from unseen test set
- Single observation!

Input pose

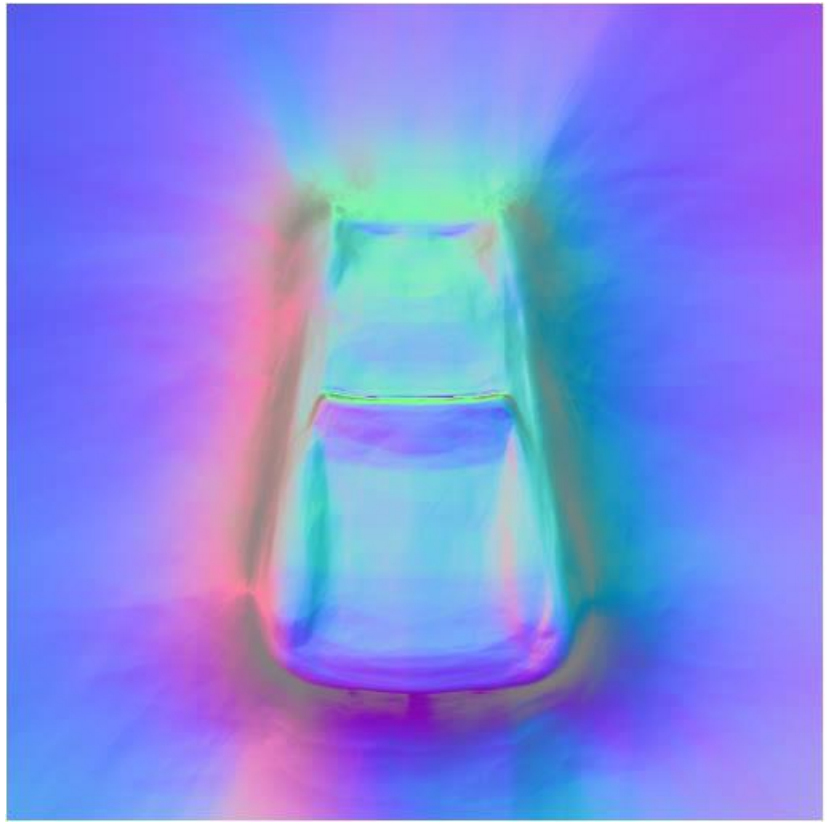
Novel View Synthesis – SRN Output

Shapenet v2 – *single-shot reconstruction* of objects in held-out test set

Input pose

Sampling at arbitrary resolutions

512x512



Surface Normals

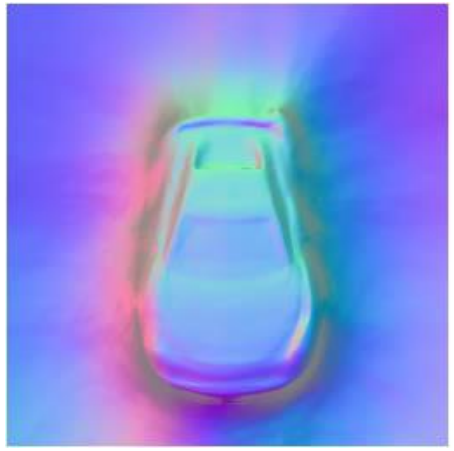
RGB

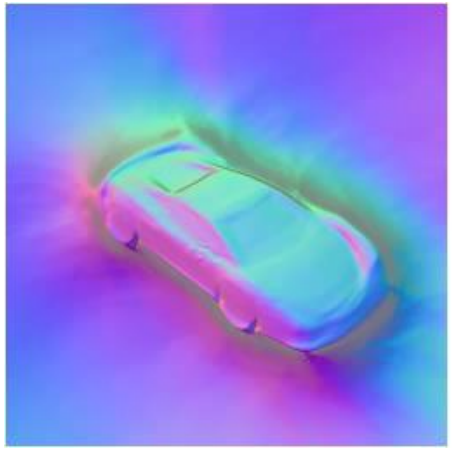
Generalization to unseen camera poses

Camera close-up

Camera Roll

SRNs



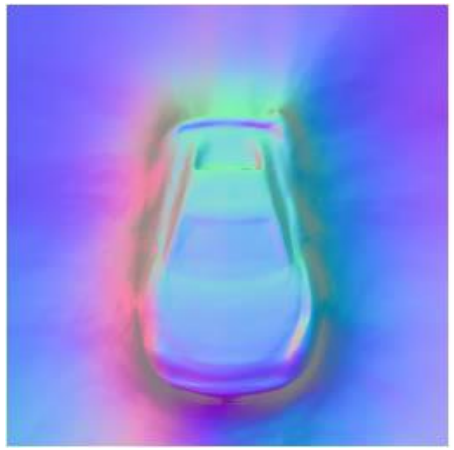


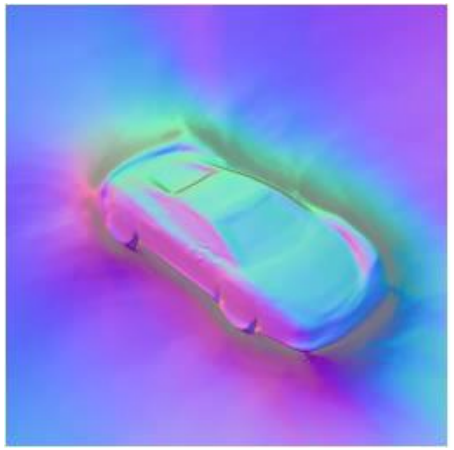
Generalization to unseen camera poses

Camera close-up

Camera Roll

SRNs



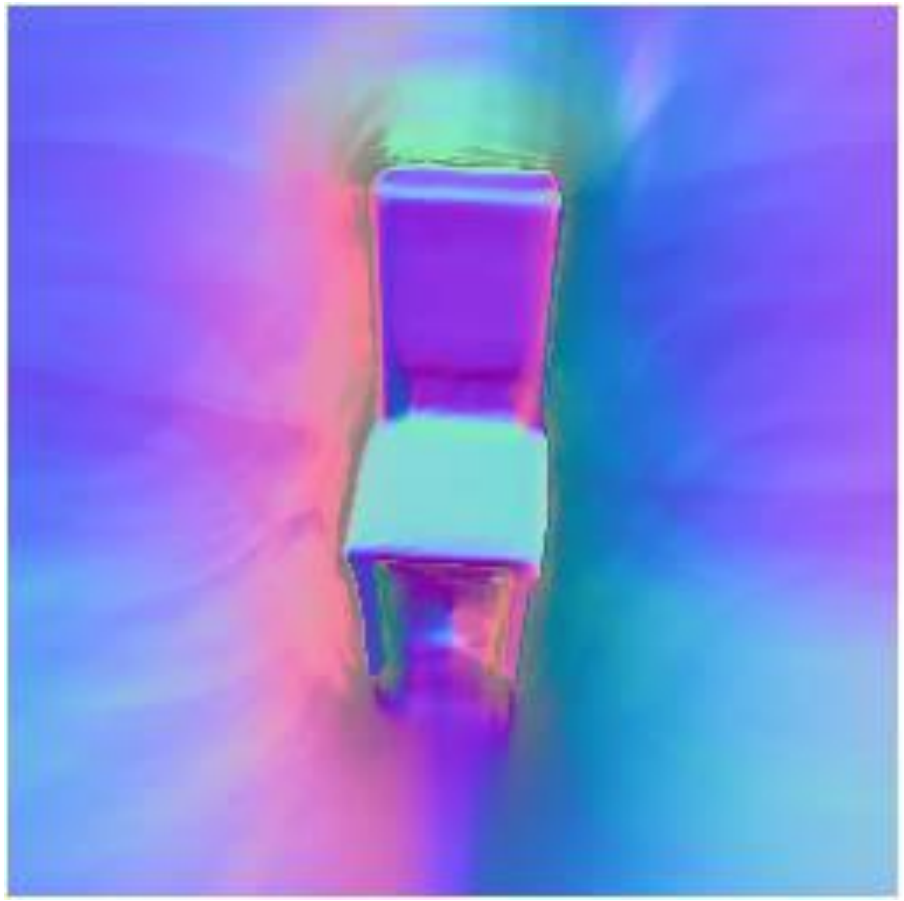


Tatarchenko et al.

Doesn't reconstruct geometry

Doesn't reconstruct geometry

Latent code interpolation



Surface Normals

RGB

Latent code interpolation

Surface Normals

RGB

Scene Representation Networks: Continuous 3D-structure-aware Neural Scene Representations

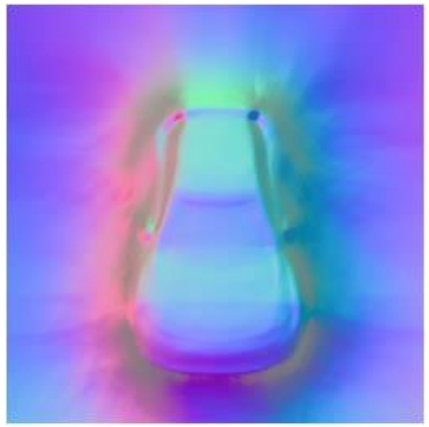
Vincent Sitzmann Michael Zollhöfer Gordon Wetzstein

Find me at Poster # 71!

Looking for research positions in scene representation learning.

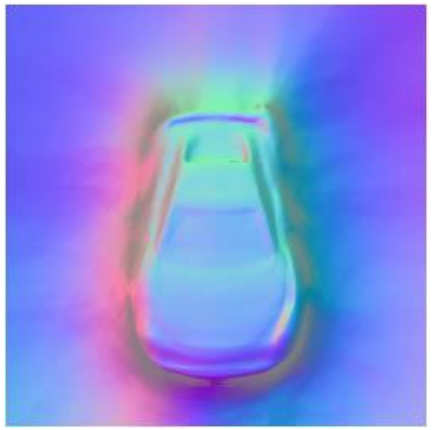
vsitzmann.github.io

@vincesitzmann



Interpolation

Single-shot reconstruction



Camera pose extrapolation