Scene Representation Learning **Without Camera Poses**

6.S980

Goal: Train AI to Infer 3D Representations from Images

From what Data?

Key signal: “Predicting the Future”

Scaling egocentric vision: The epic-kitchens dataset, Damen et al., CVPR 2018

https://www.popularmechanics.com/technology/robots/a20190/google-room-full-of-robot-arms/

The Structure of Video

≈

3D Geometry & Appearance

+

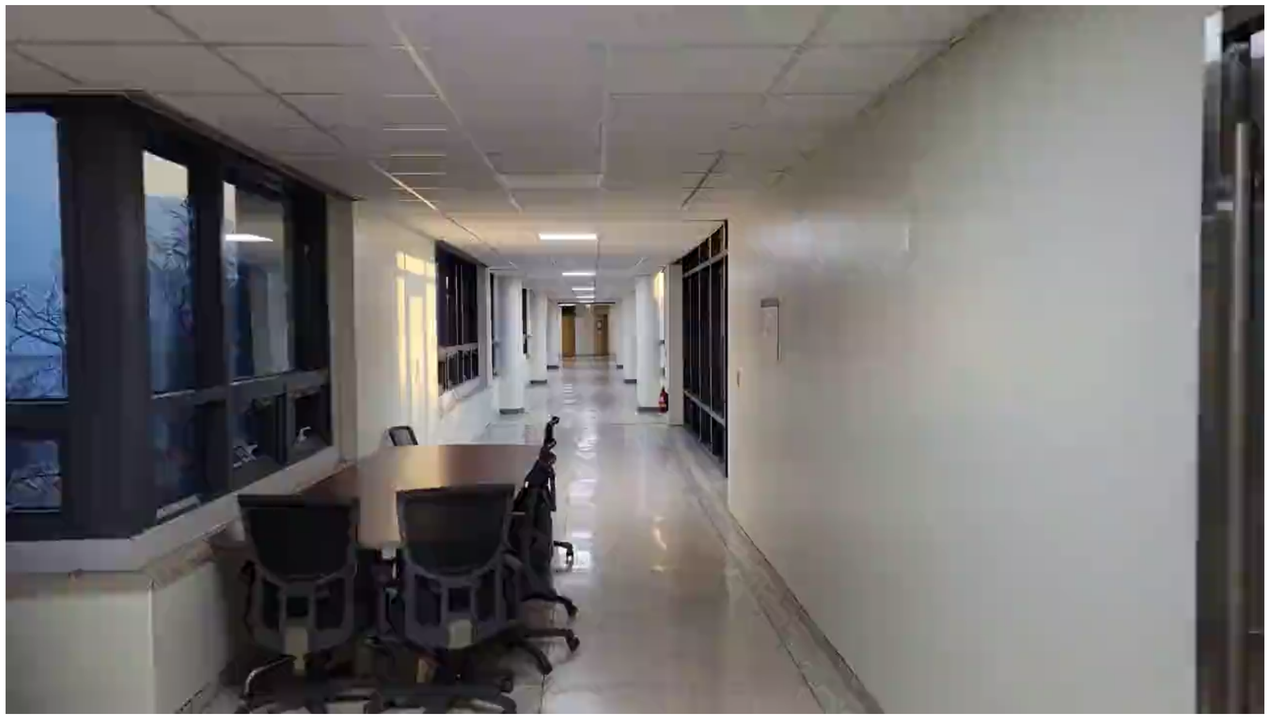
Camera / Ego-Motion

+

3D Motion & Deformation

Scaling egocentric vision: The epic-kitchens dataset, Damen et al., CVPR 2018

The Structure of Video



≈

3D Geometry & Appearance

+

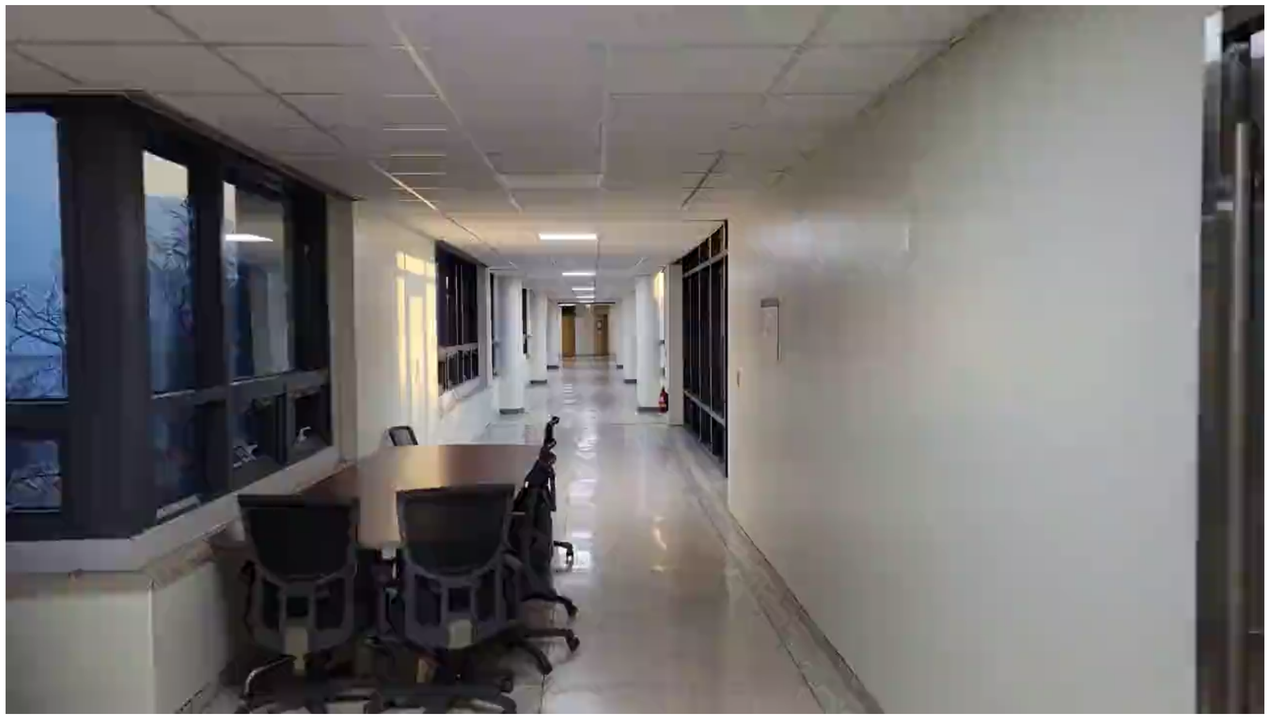
Camera / Ego-Motion

+

3D Motion & Deformation

Progressively Optimized Local Radiance Fields for Robust View Synthesis, Meuleman et al., CVPR 2023

The Structure of Video



≈

3D Geometry & Appearance

Assume given +

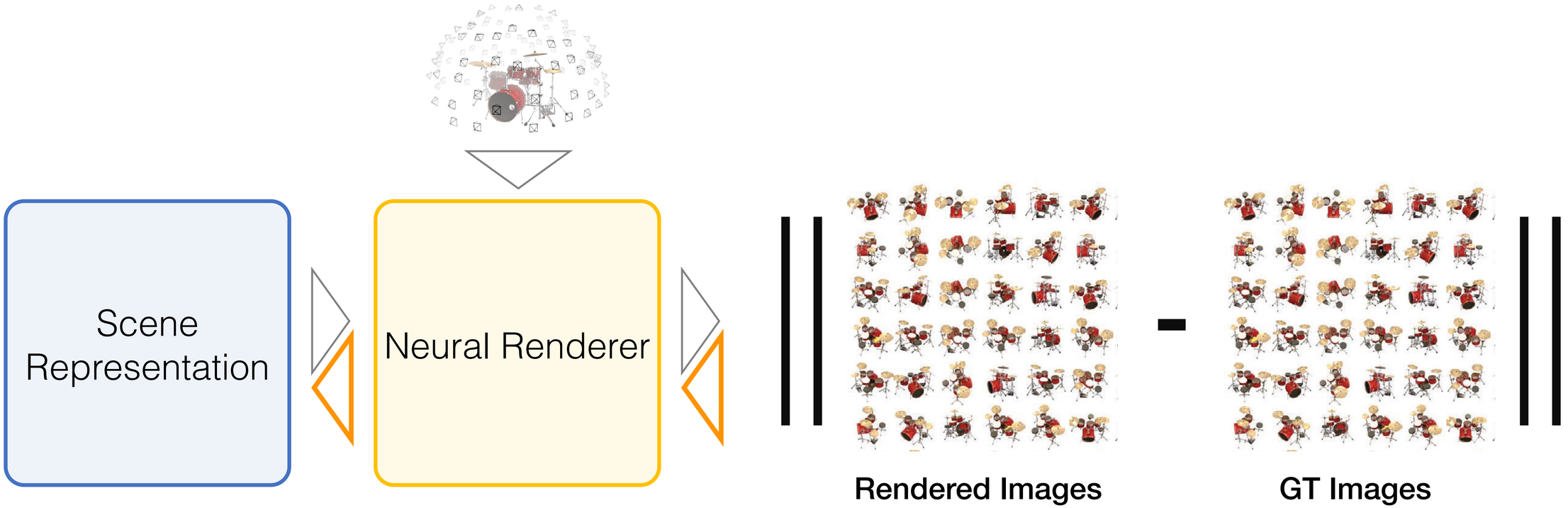
Camera / Ego-Motion

+

3D Motion & Deformation

Progressively Optimized Local Radiance Fields for Robust View Synthesis, Meuleman et al., CVPR 2023

3D Reconstruction from *hundreds* of Images



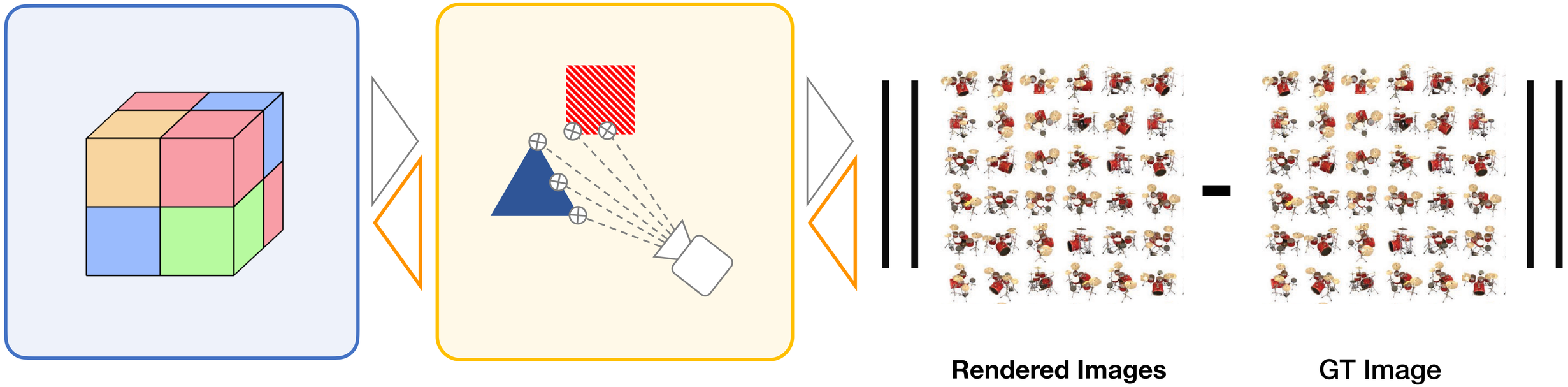
3D Reconstruction from *hundreds* of Images

Is this a candidate for vision foundation models?

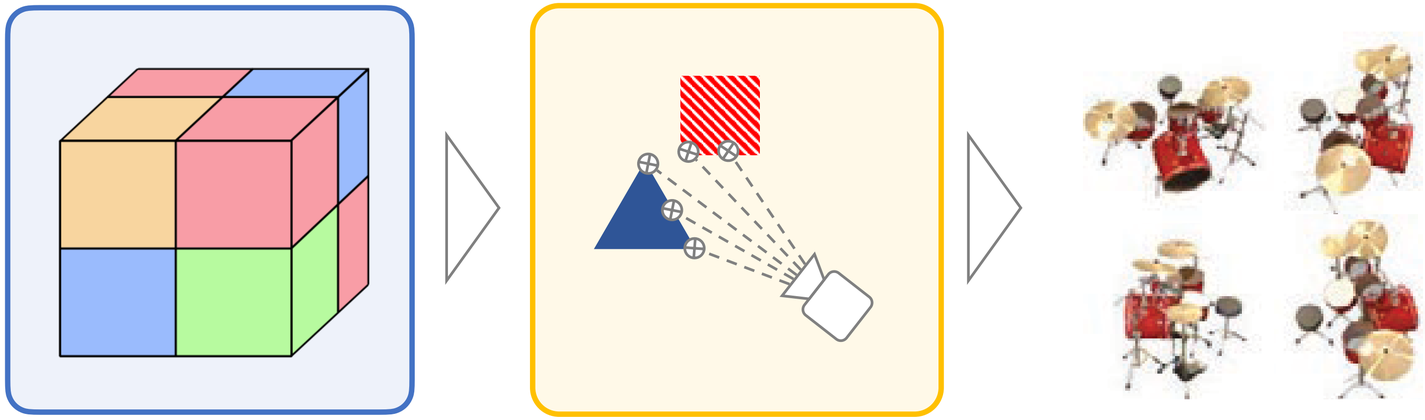
NeRF, ECCV 2021

Plenoxels, CVPR 2022

Not scalable: Have to reconstruct each scene separately from *hundreds* of images

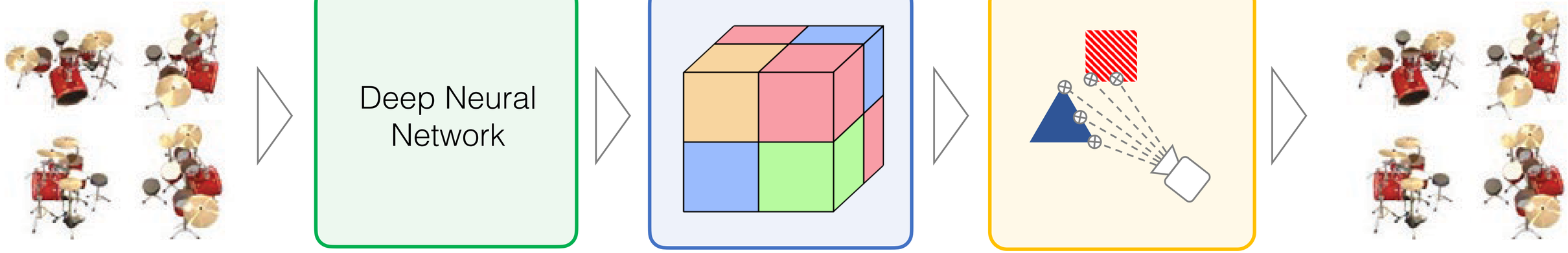


Not scalable: Have to reconstruct each scene separately…

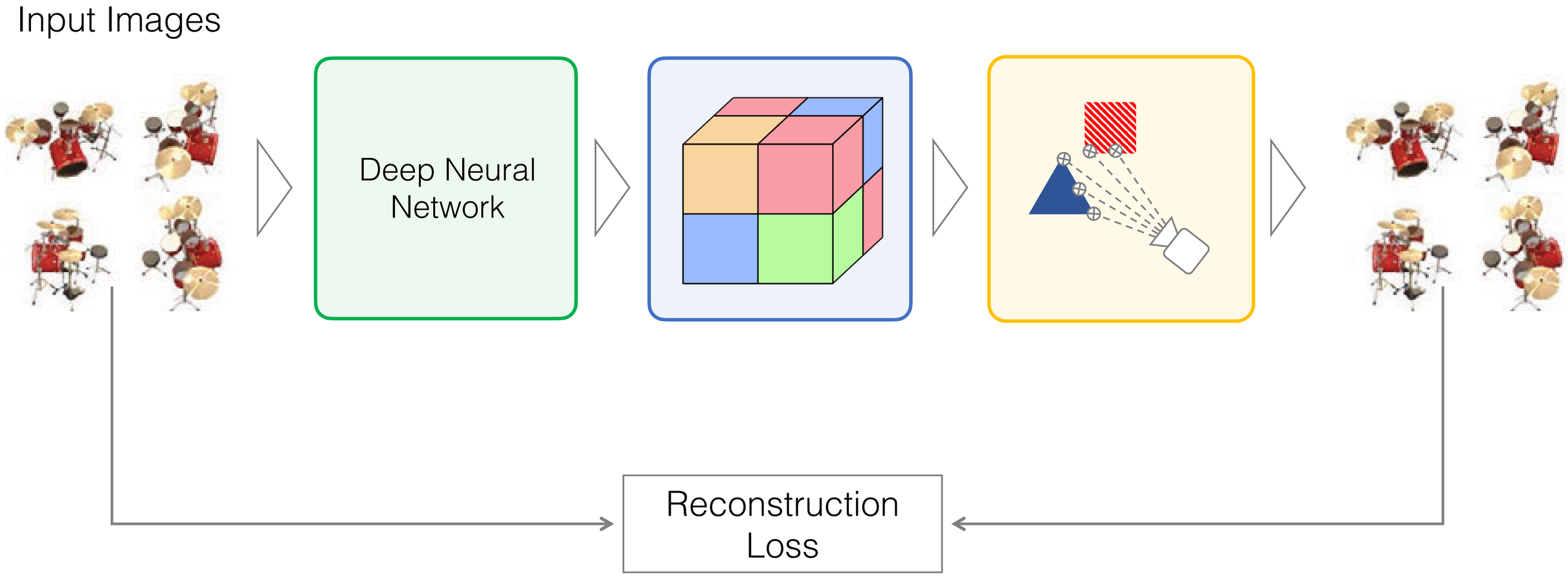


Amortized (=feedforward, generalizable) 3D Reconstruction

Input Images



Amortized (=feedforward, generalizable) 3D Reconstruction



Large-baseline novel view synthesis from **two images**

*Is **this** a candidate for a "vision foundation model"?*

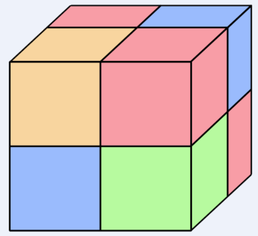
Learning to Render Novel Views from Wide-Baseline Stereo Pairs, Du et al., CVPR 2023

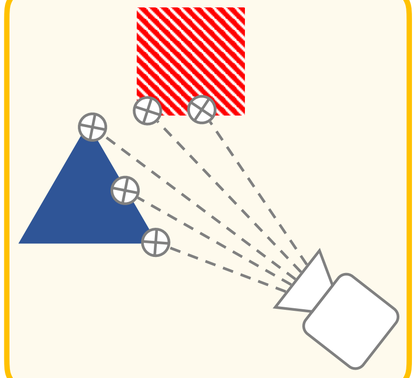
Amortized (=feedforward, generalizable) 3D Reconstruction

Input Images

Deep Neural Network

$\Phi : \mathbb{R}^3 \rightarrow \mathbb{R}^n$

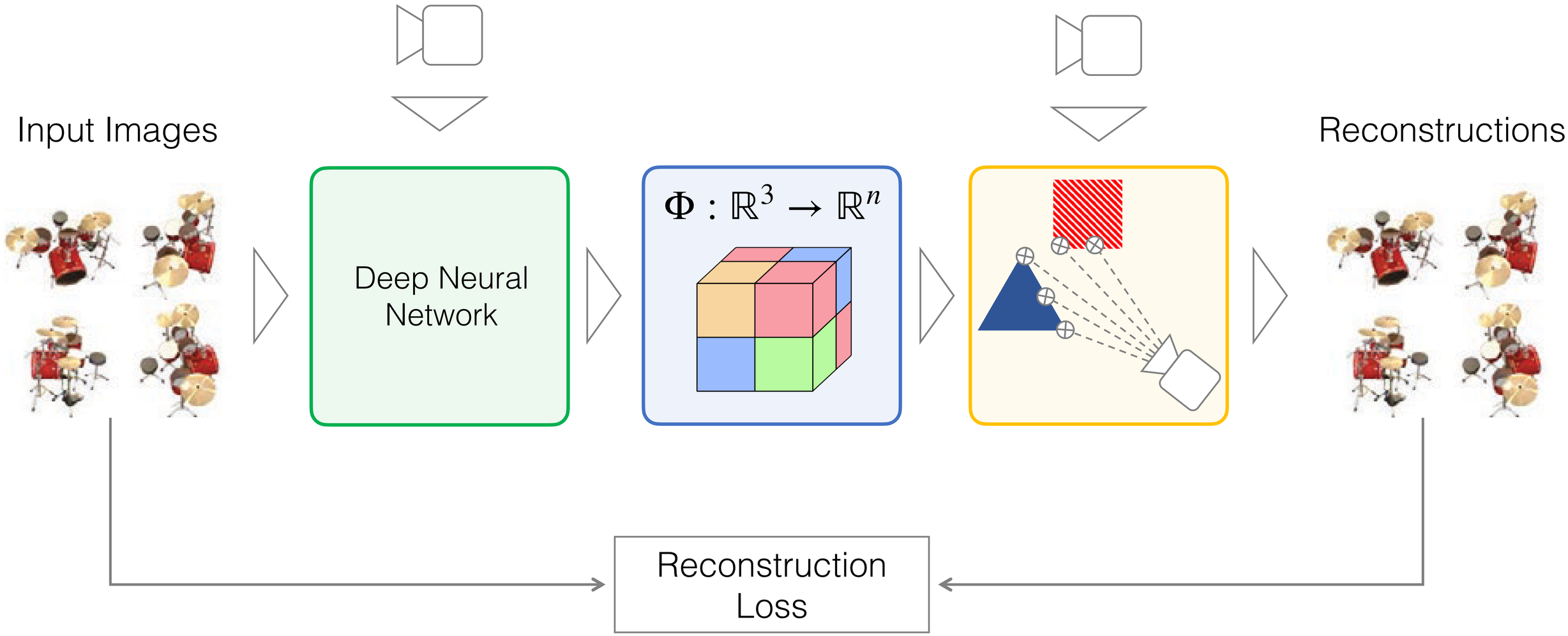




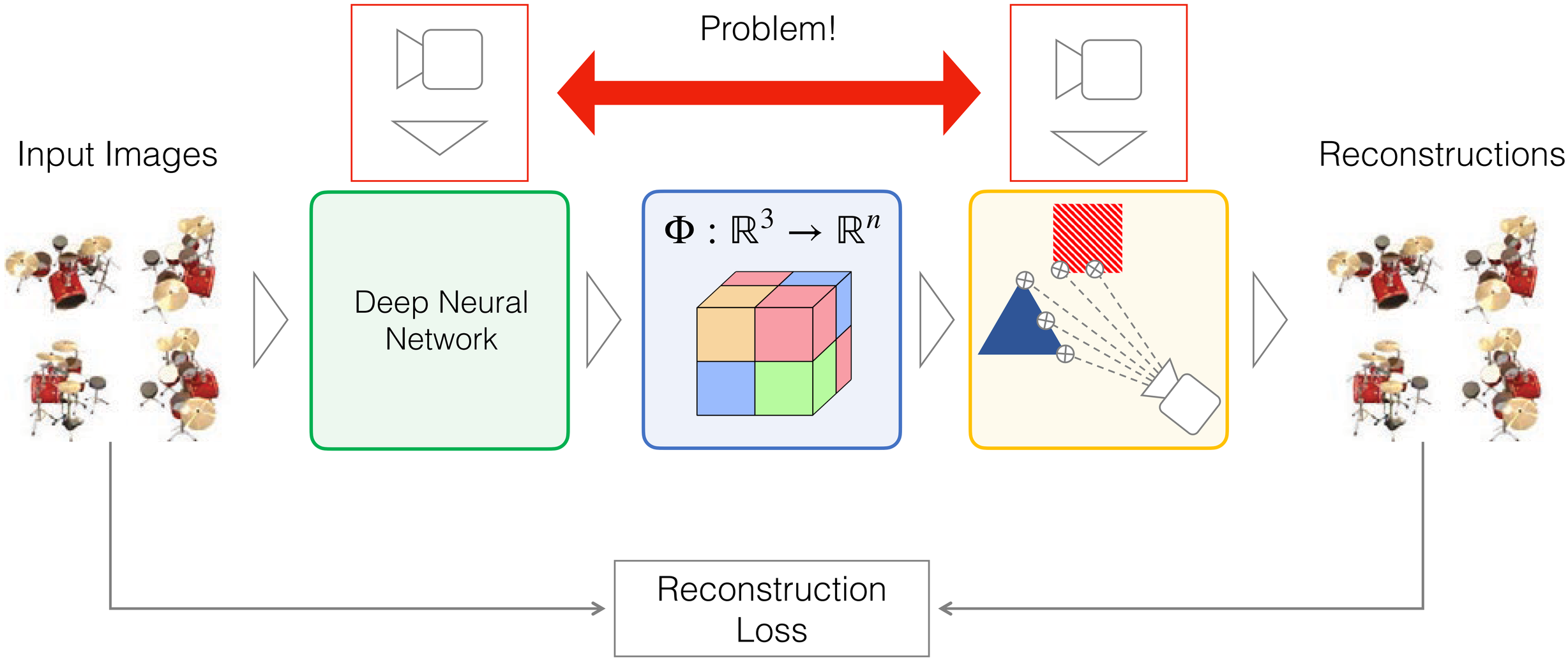
Reconstructions

Reconstruction Loss

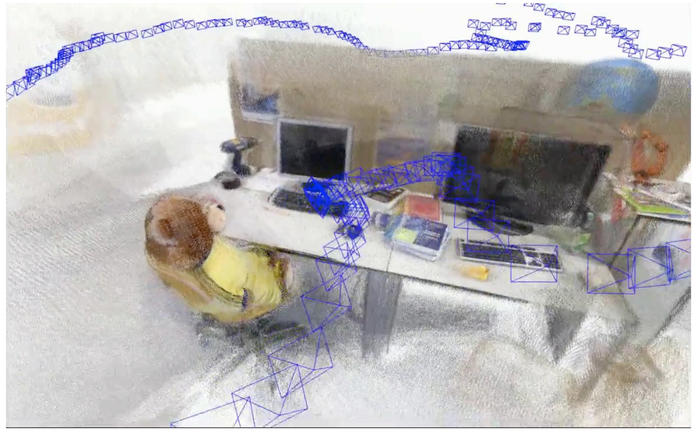
Amortized (=feedforward, generalizable) 3D Reconstruction



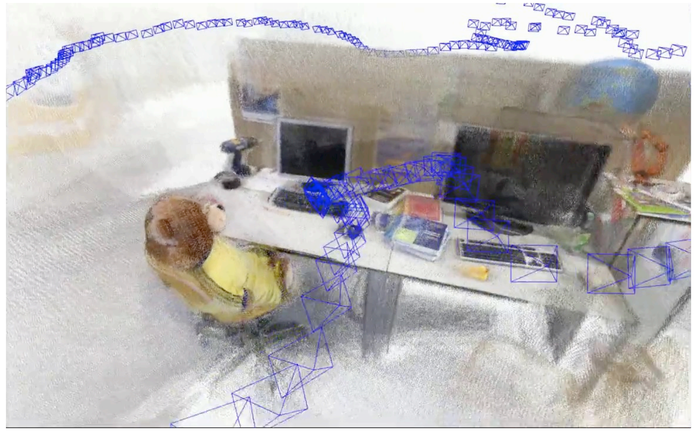
Amortized (=feedforward, generalizable) 3D Reconstruction

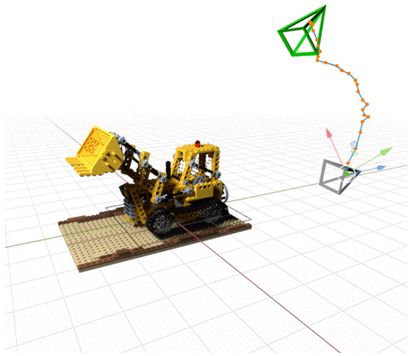


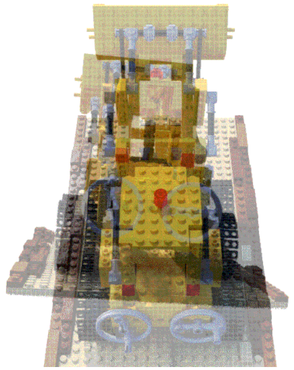
Existing Approaches



Existing Approaches







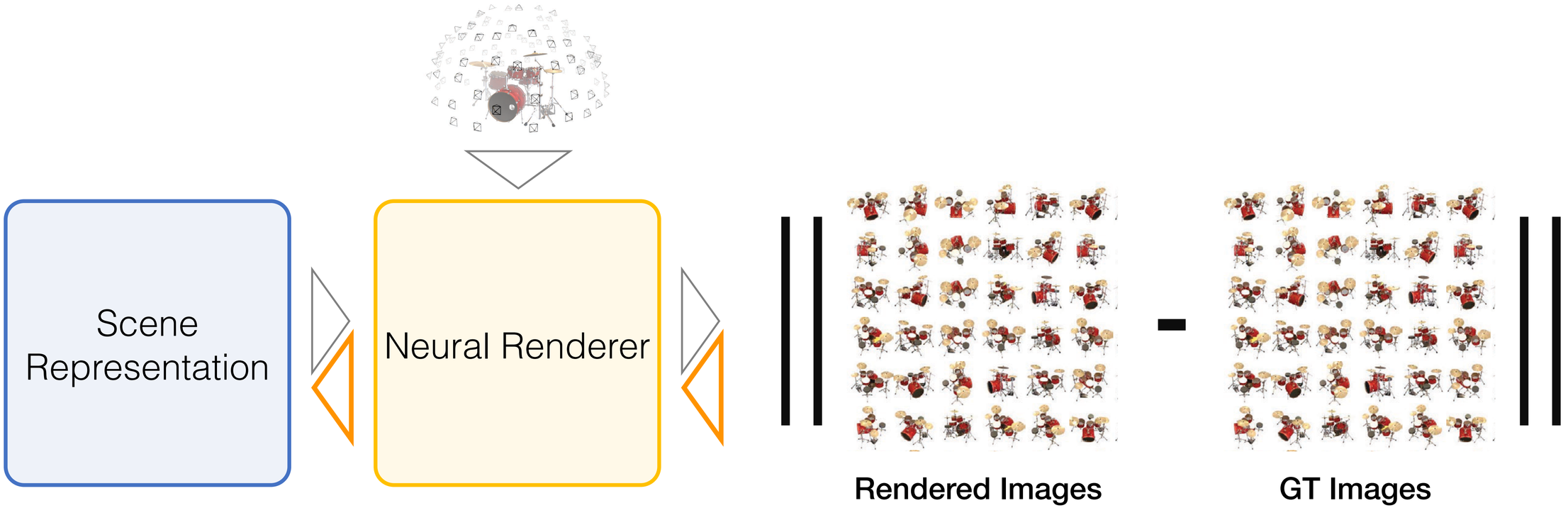
SFM & SLAM

- COLMAP
- ORB-SLAM
- DROID-SLAM
- …

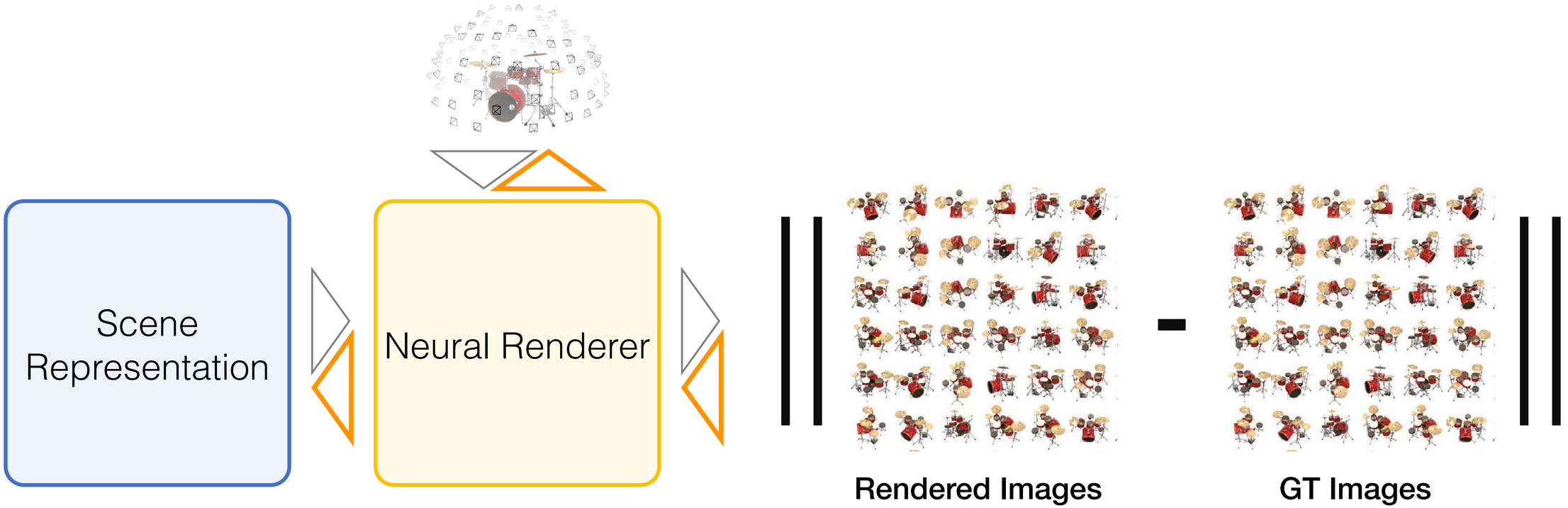
Joint NeRF & pose optim.

- iNeRF, Yen-Chen et al.
- BARF, Lin et al.
- MELON, Levy et al.
- …

3D Reconstruction from *hundreds* of Images



3D Reconstruction from *hundreds* of Images

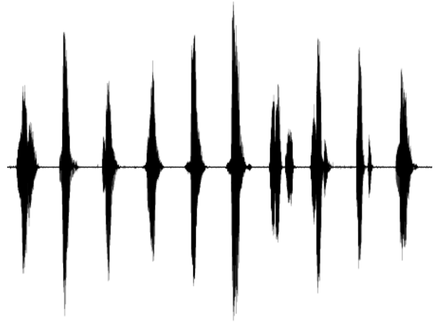


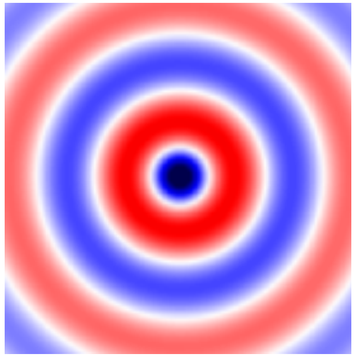
Images

Shapes

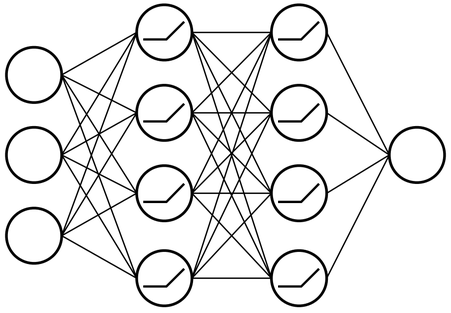
Audio

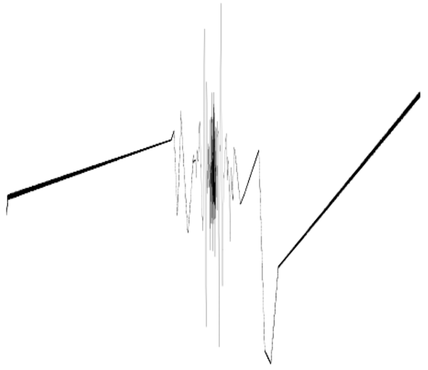
Quantities defined by a differential equation

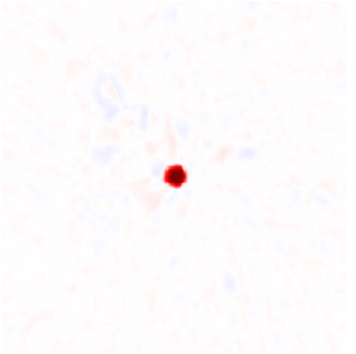




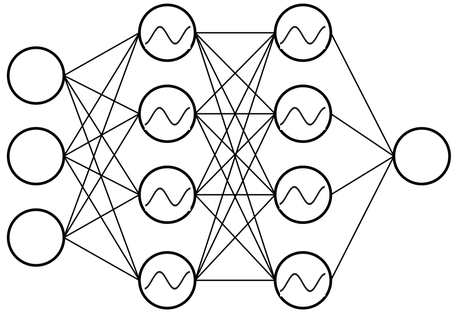
ReLU MLP

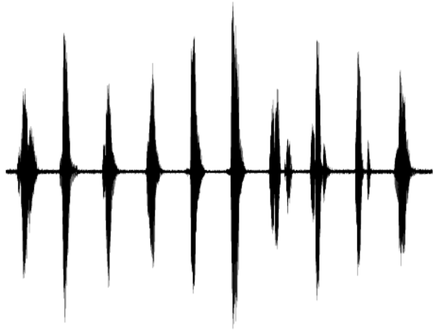


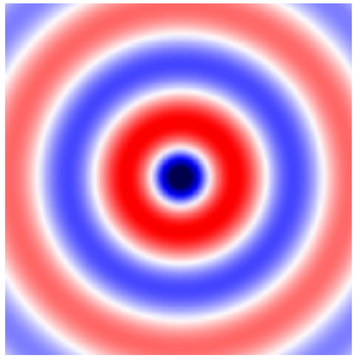




SIREN



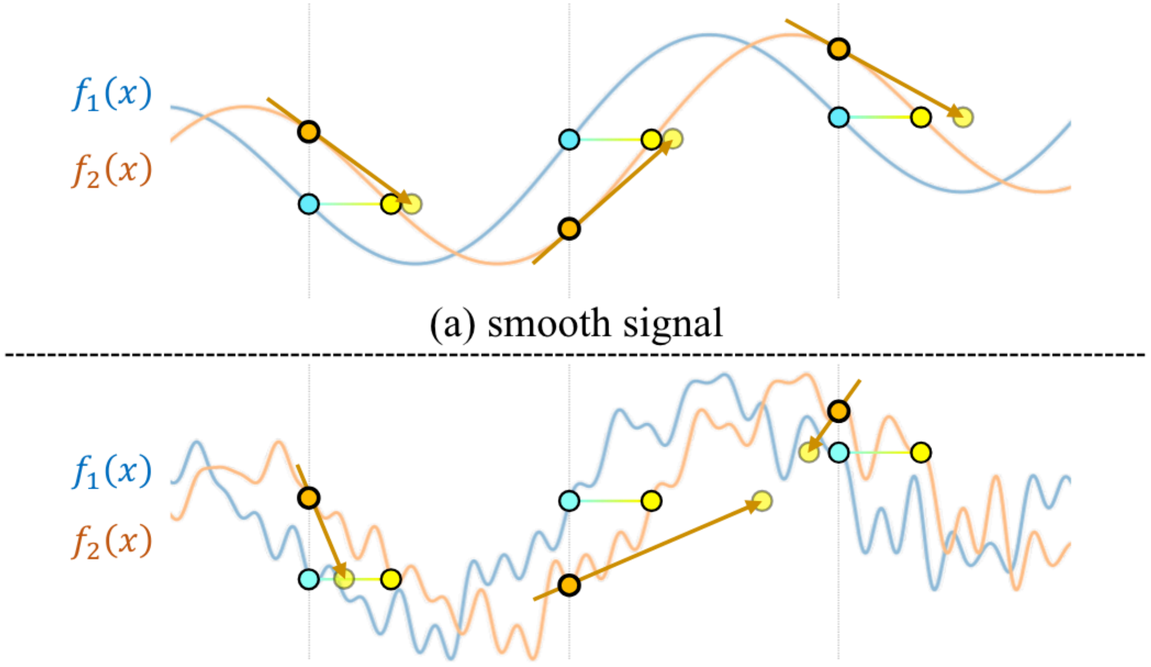




BARF 🤮: Bundle-Adjusting Neural Radiance Fields

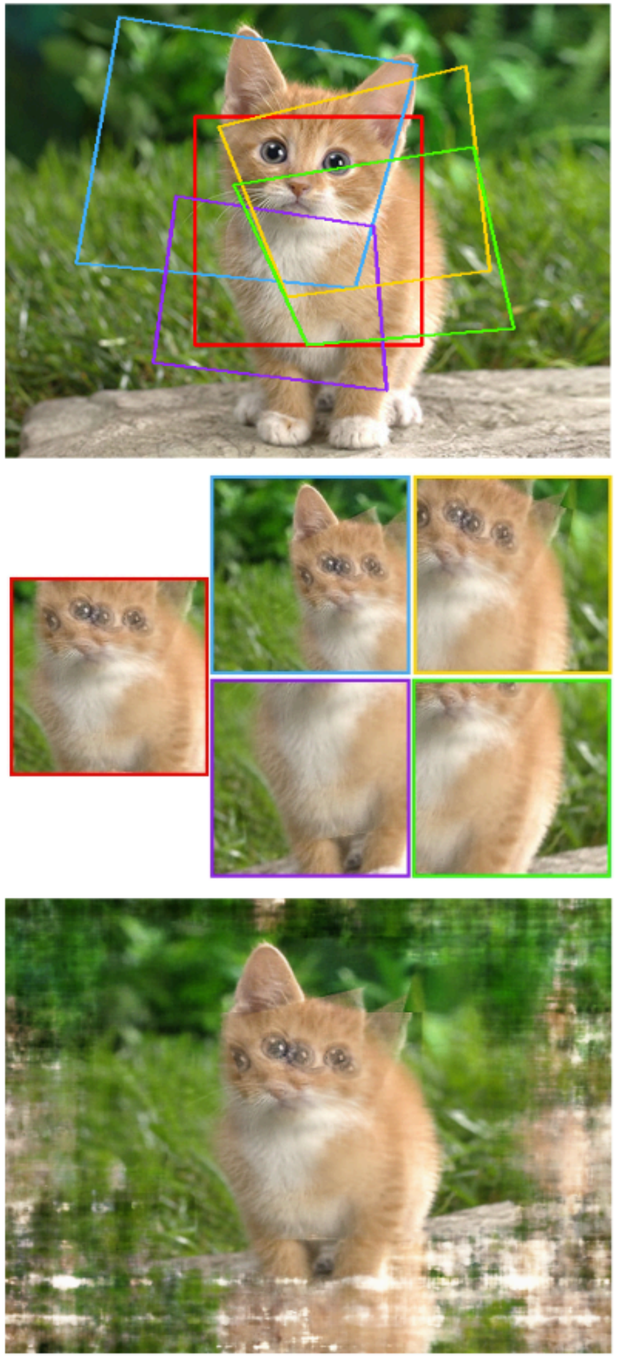
Chen-Hsuan Lin[1] Wei-Chiu Ma[2] Antonio Torralba[2] Simon Lucey[1,3]
[1]Carnegie Mellon University [2]Massachusetts Institute of Technology [3]The University of Adelaide

https://chenhsuanlin.bitbucket.io/bundle-adjusting-NeRF

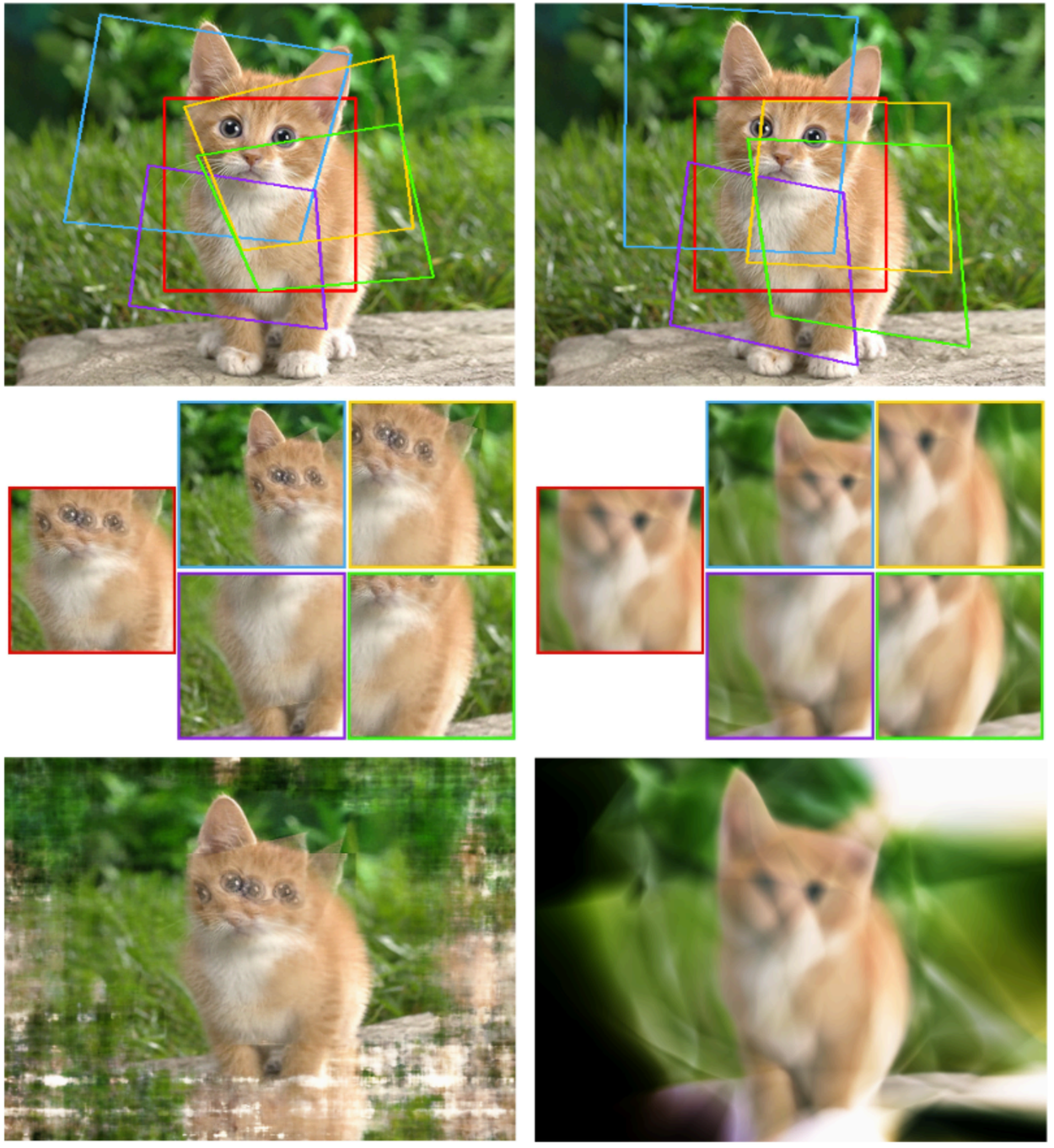


(a) smooth signal

(b) complex signal



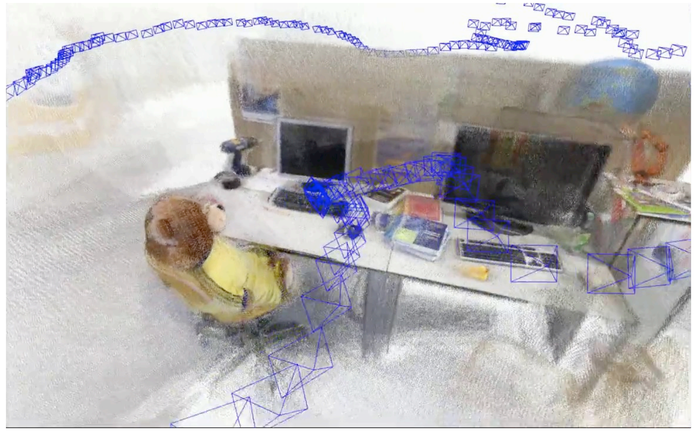
(a) naïve pos. enc.

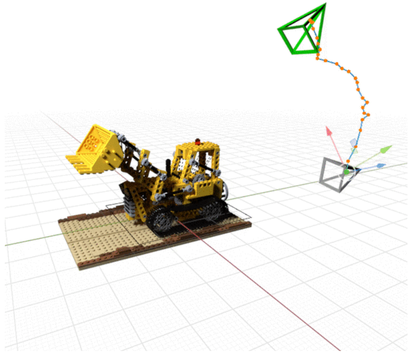


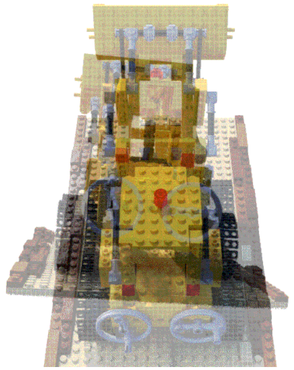
(a) naïve pos. enc. (b) w/o pos. enc.

(a) naïve pos. enc. (b) w/o pos. enc. (c) BARF

Existing Approaches







SFM & SLAM

- COLMAP
- ORB-SLAM
- DROID-SLAM
- …

Joint NeRF & pose optim.

- iNeRF, Yen-Chen et al.
- BARF, Lin et al.
- MELON, Levy et al.
- …

Joint depth & pose pred.

- Monodepth, Godard et al.
- Unsupervised Depth and Ego-Motion…, Zhou et al.
- Self-supervised Learning of Depth and…, Guizilini et al.

Unsupervised Depth Prediction!

Frame at time t_1

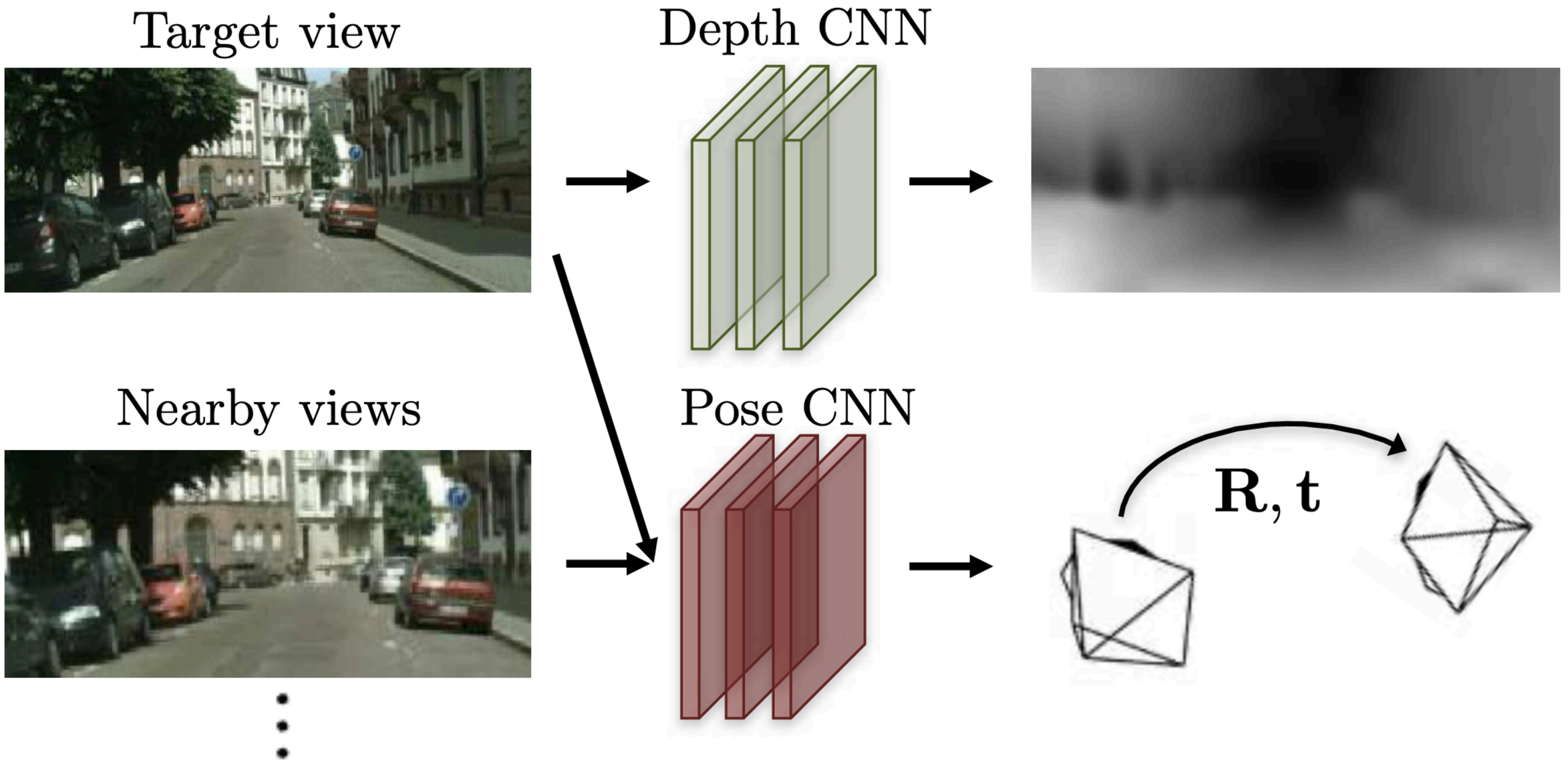
...

Frame at time t_2

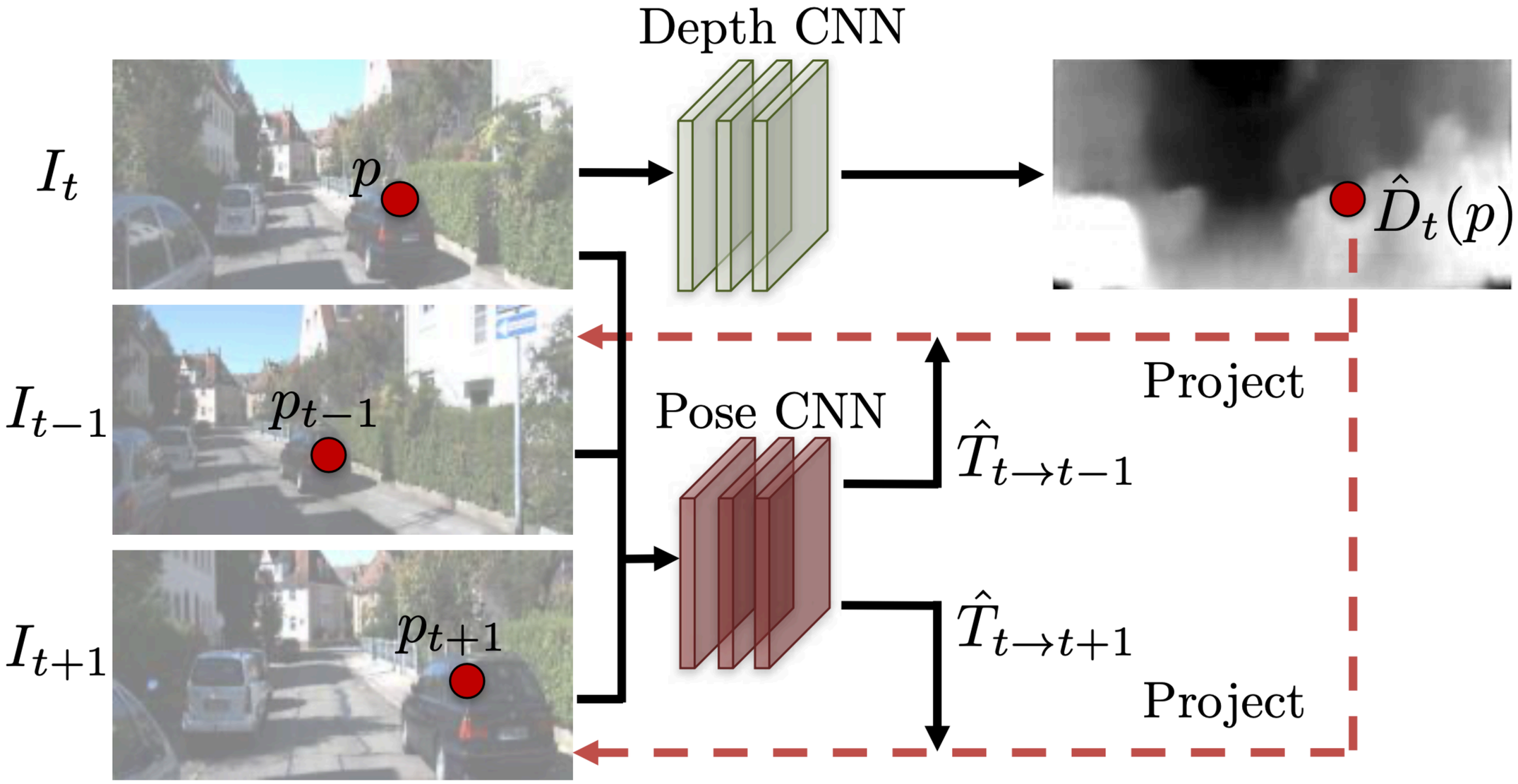
Goal: Learn Depth and Ego-Motion (relative camera pose) just from video!

How?

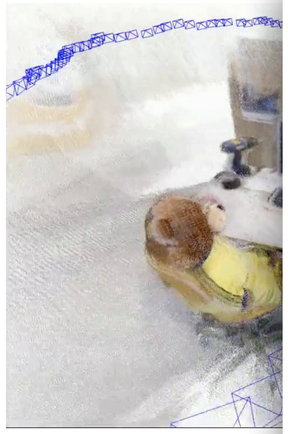
Unsupervised Depth and Ego-Motion from Video (Zhou et al. 2017)



Unsupervised Depth and Ego-Motion from Video (Zhou et al. 2017)



Conventional Approaches



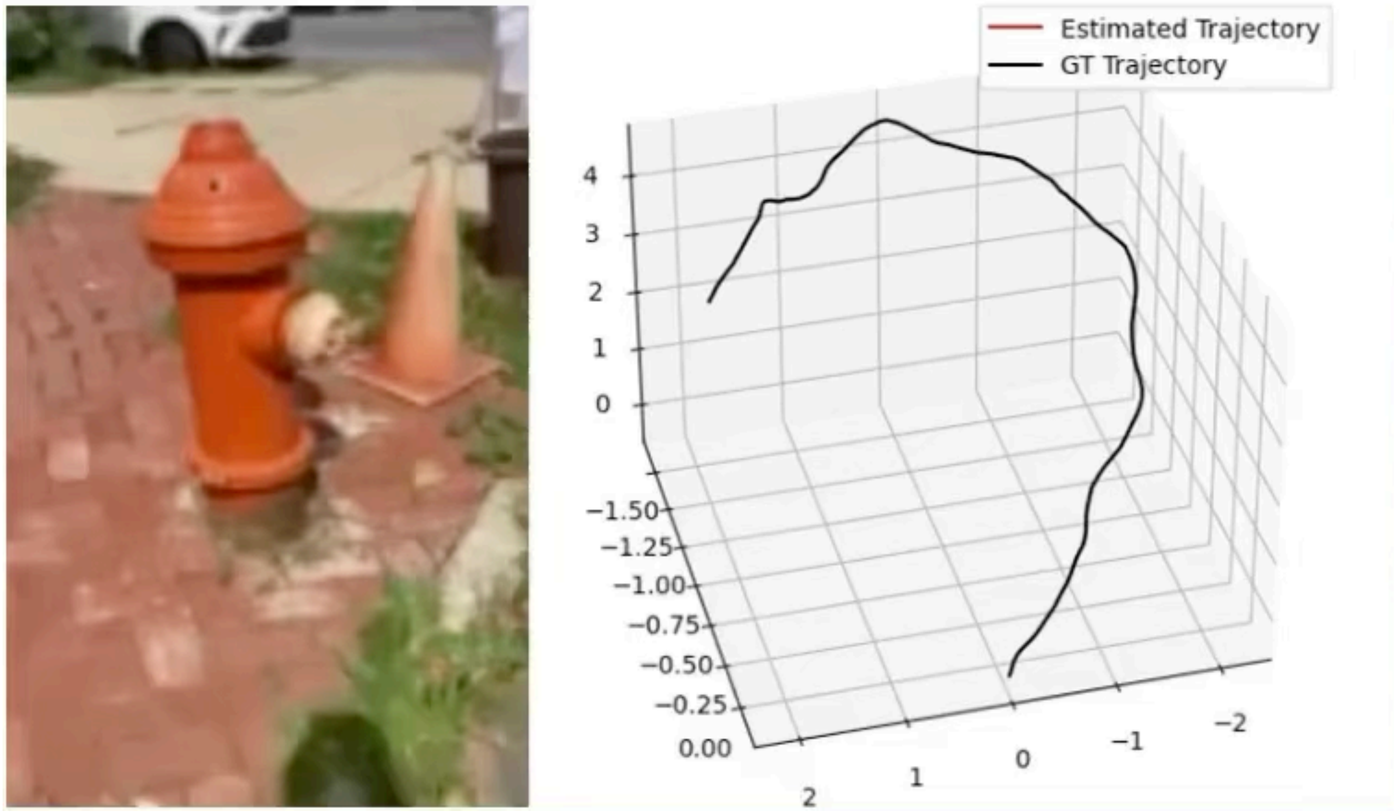
In practice: The only area of computer vision
where the state-of-the-art does not use learned priors.

SFM & SL **pred.**

- Per-sce
- If using
- SLAM: Prone to failure for particular trajectories, doesn't estimate camera intrinsics!

FlowCam:

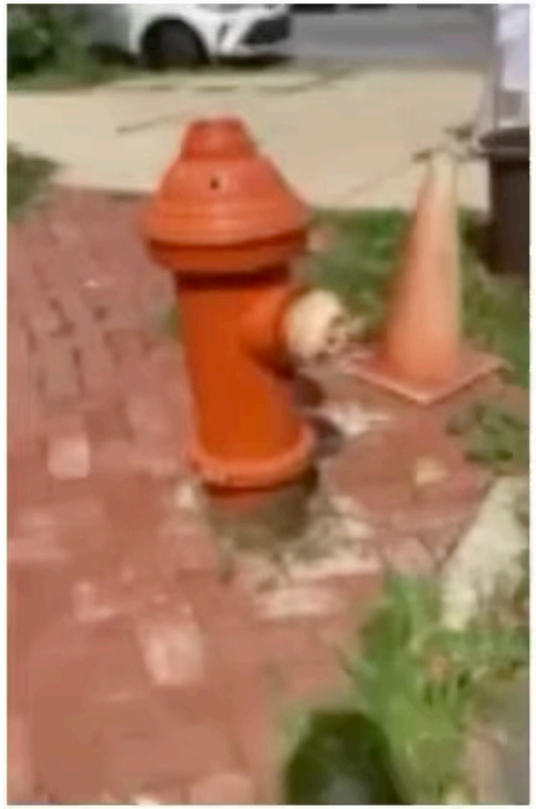
Generalizable Neural Scene Representations without Camera Poses

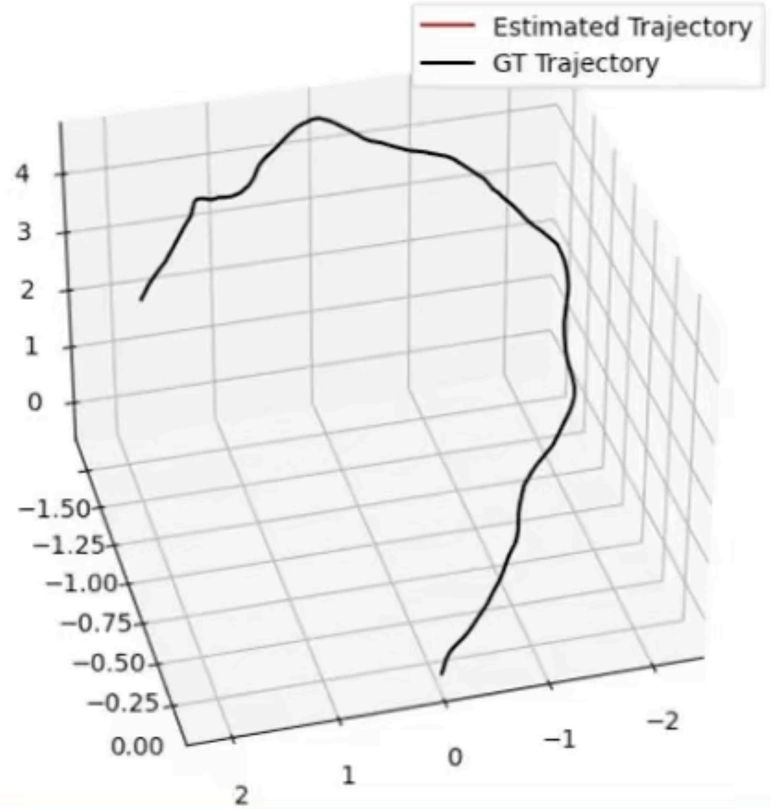


Poses inferred **feed-forward at ~20FPS. No scene-specific optimization.**

FlowCam:

Generalizable Neural Scene Representations without Camera Poses





Renderings from Generalizable Neural Scene Representation.

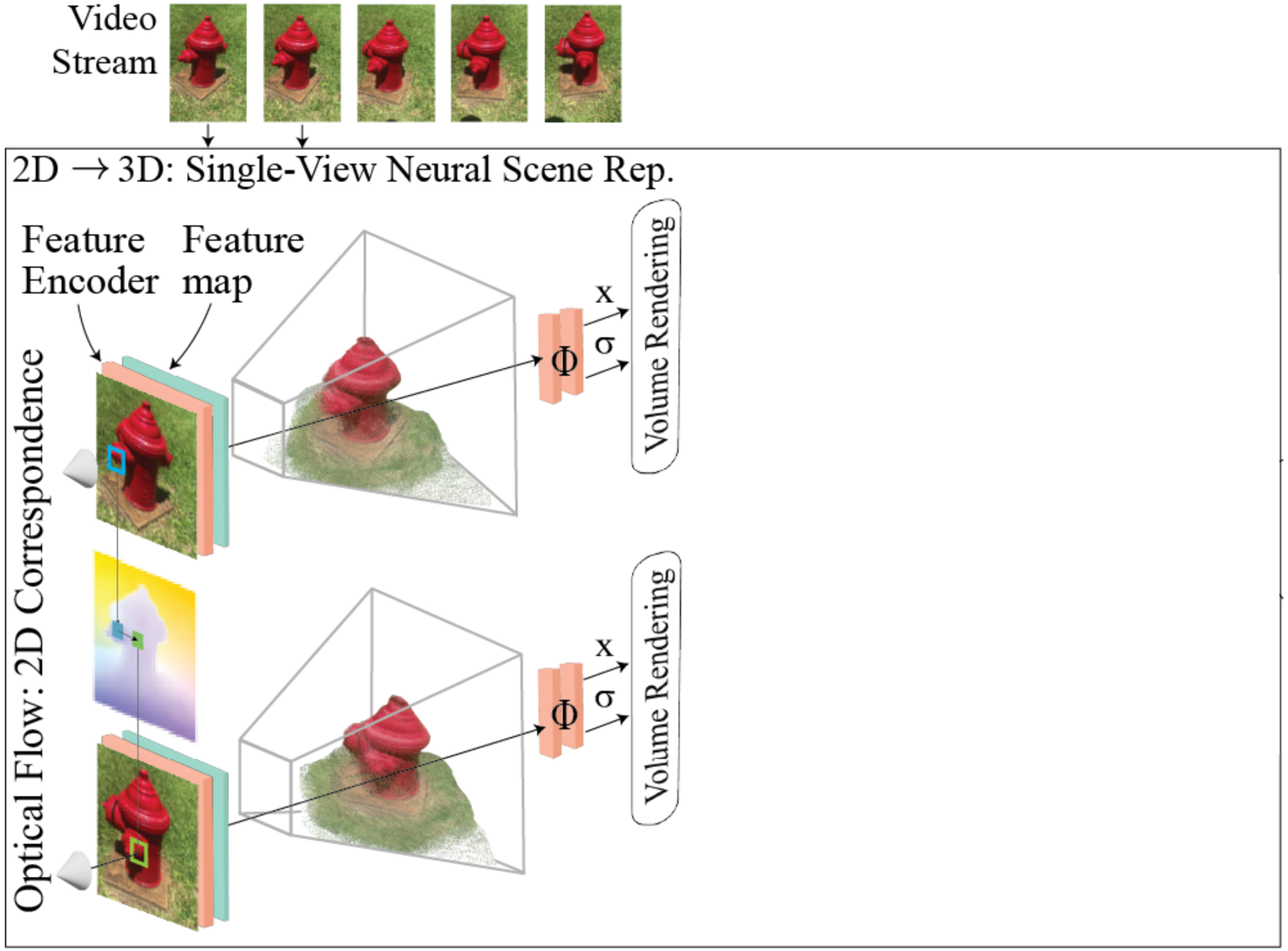
Trained end-to-end, without camera poses, only on raw video!

Step 1: Differentiable Pose Estimation

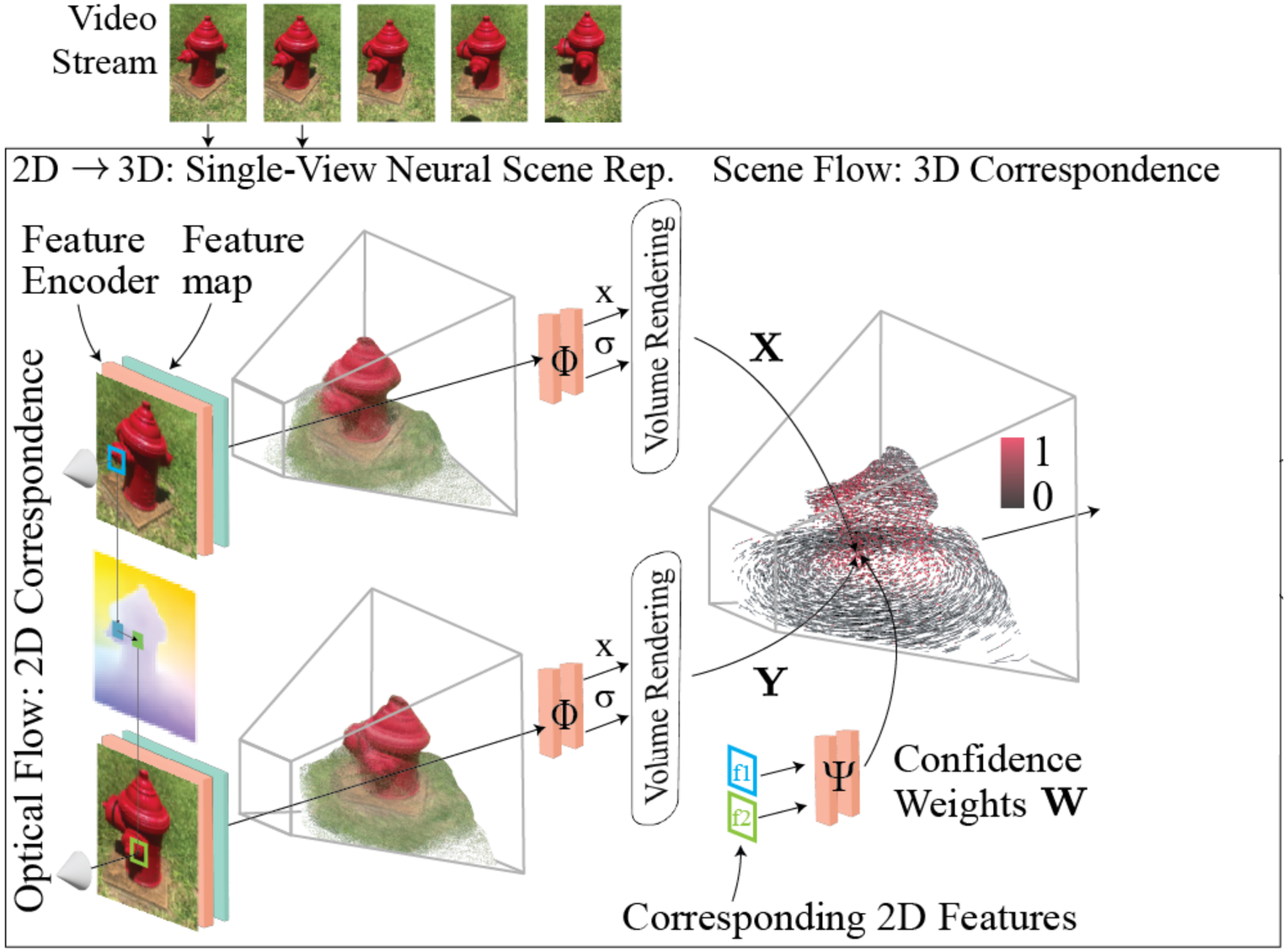
Video Stream

Step 1: Differentiable Pose Estimation

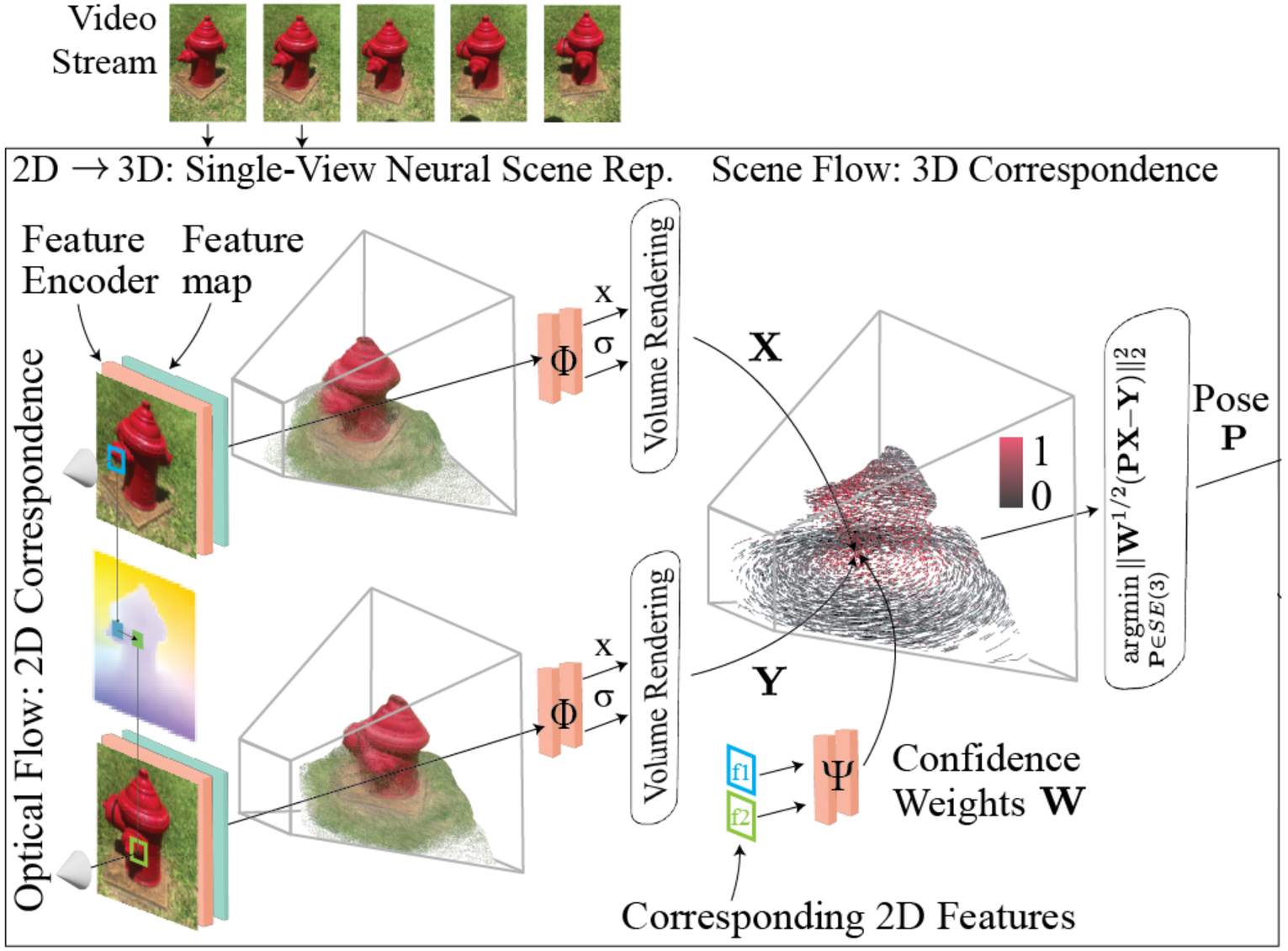
Step 1: Differentiable Pose Estimation



Step 1: Differentiable Pose Estimation

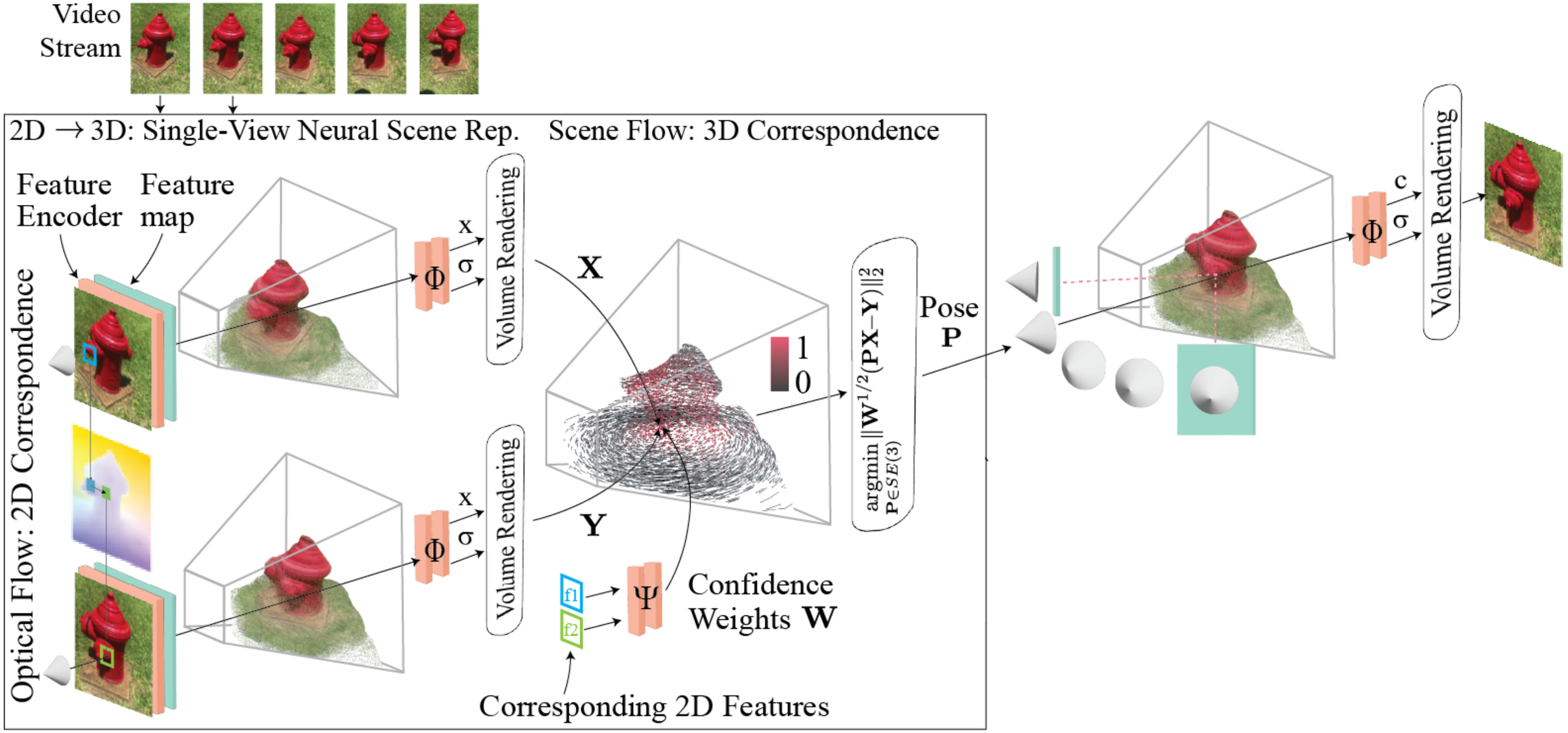


Step 1: Differentiable Pose Estimation



Step 1: Differentiable Pose Estimation

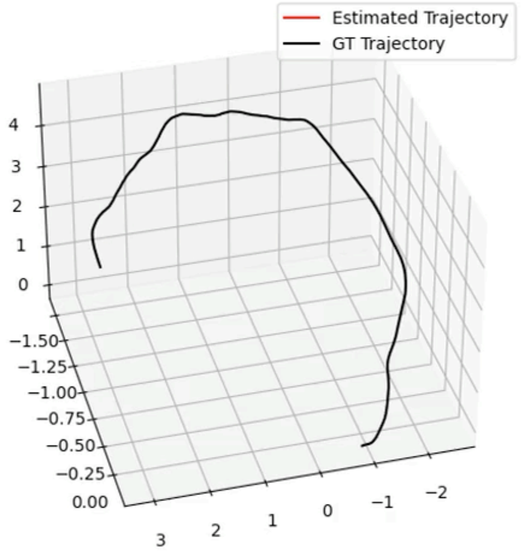
Step 2: Video Re-Rendering



Co3D - Results

Input

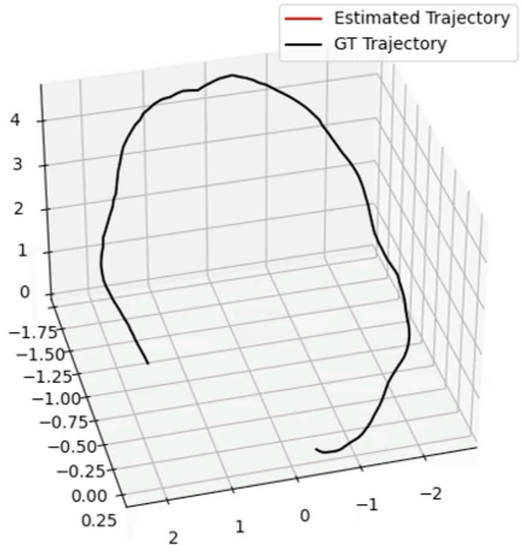
Est. Poses



Novel Views

Input

Est. Poses

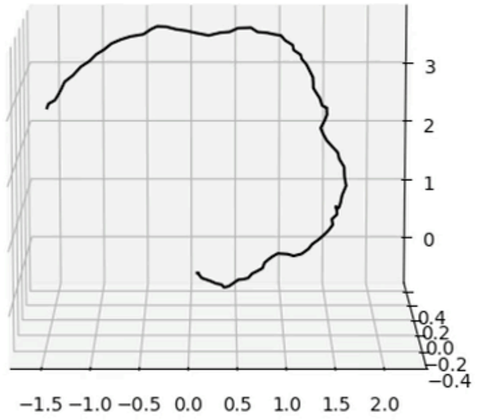


Novel Views

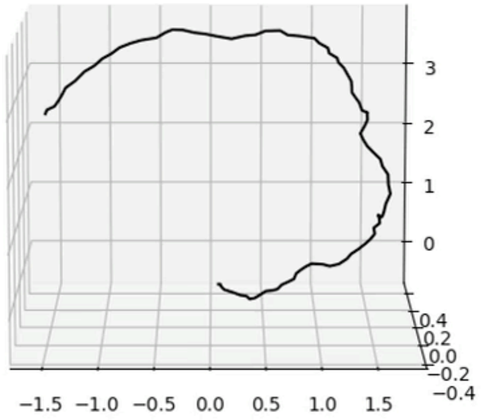
Result: Tanks & Temples

Input Video

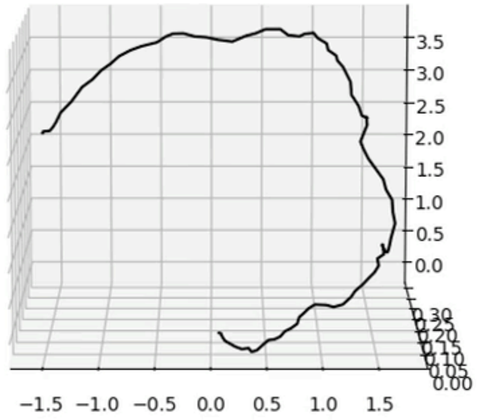
BARF Poses



Our RealEstate10k Model



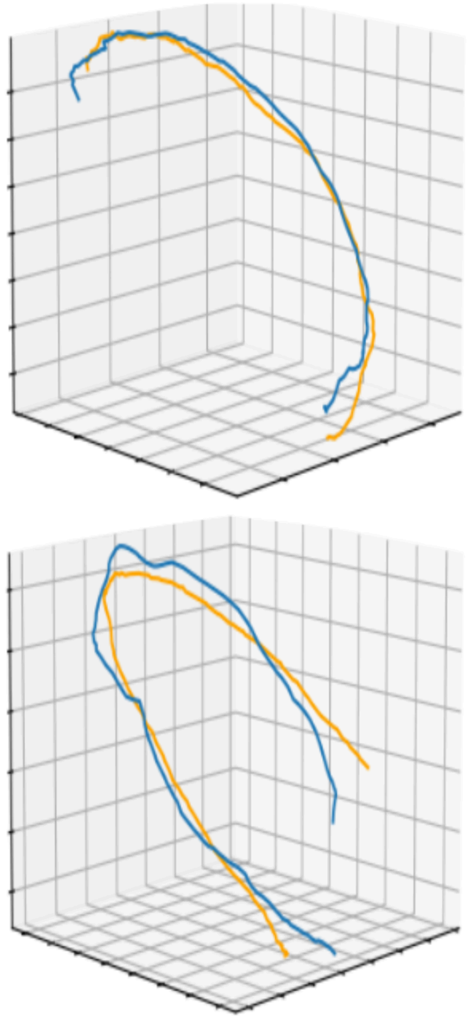
Ours Fine-Tuned



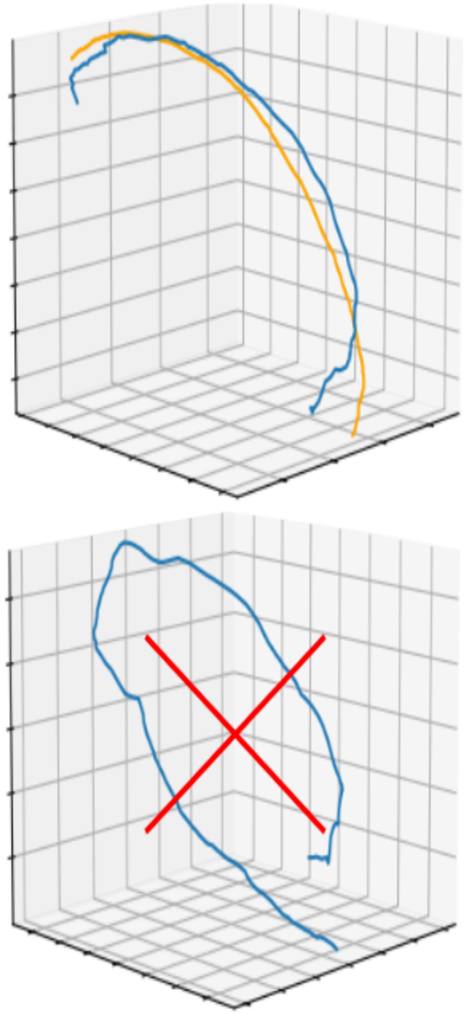
Comparison - Co3D Pose Estimation

Sample Frames

Our Poses



ORB-SLAM v3 Poses



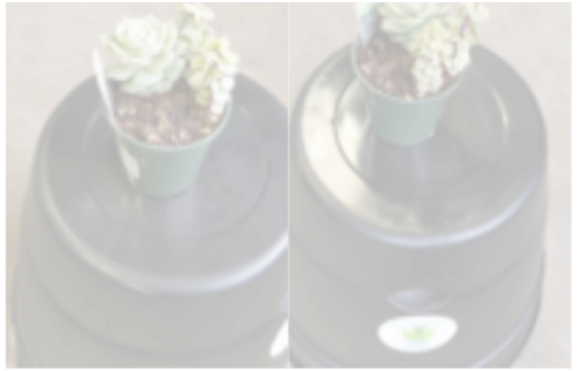
Comparison - Co3D Pose Estimation

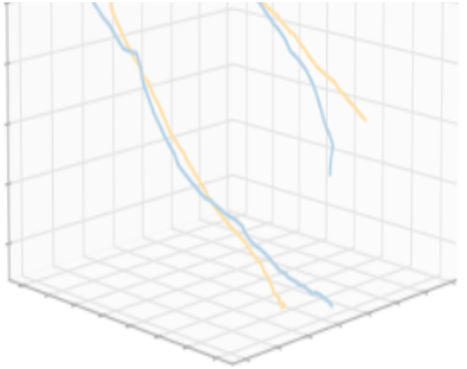
Sample Frames

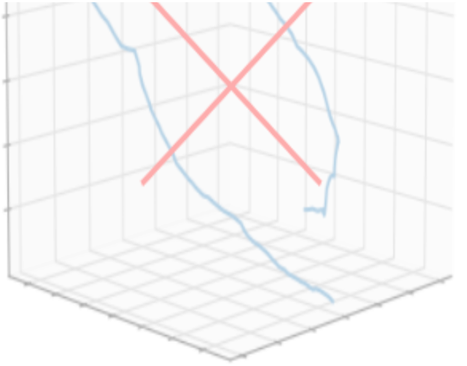
Our Poses

ORB-SLAM v3 Poses

	Top	Bot.	% Tracked
ORB3 [12]	0.28	0.69	49
DROID [33]	0.38	0.82	**100**
Ours	**0.19**	**0.32**	**100**







Limitations

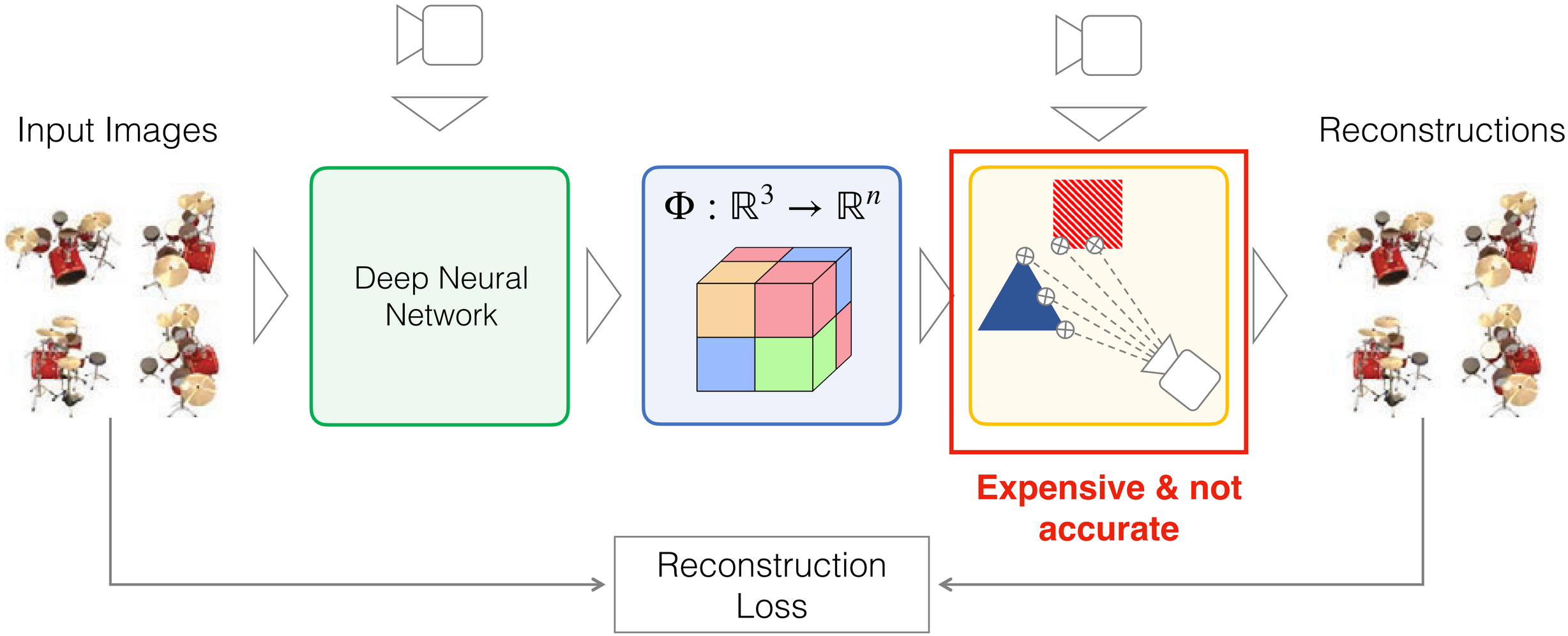
- We use a simple method to predict intrinsics, but it’s not bulletproof.
- No loop closures yet, no refinement - feed-forward it is.
- Not photo-realistic.
  Improvements in generalizable neural scene representations will help!
- Expensive training: Training requires one rendering pass per frame…
  Need faster & better differentiable renderers!

FlowCam: Training Generalizable 3D Radiance Fields without Camera Poses via Pixel-Aligned Scene Flow

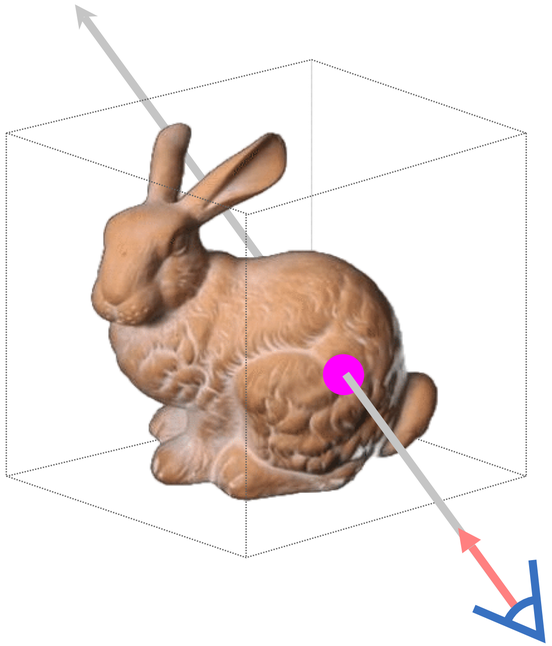
Cameron Smith **Yilun Du** **Ayush Tewari** **Vincent Sitzmann**
{camsmith, yilundu, ayusht, sitzmann}@mit.edu
MIT CSAIL
cameronosmith.github.io/flowcam

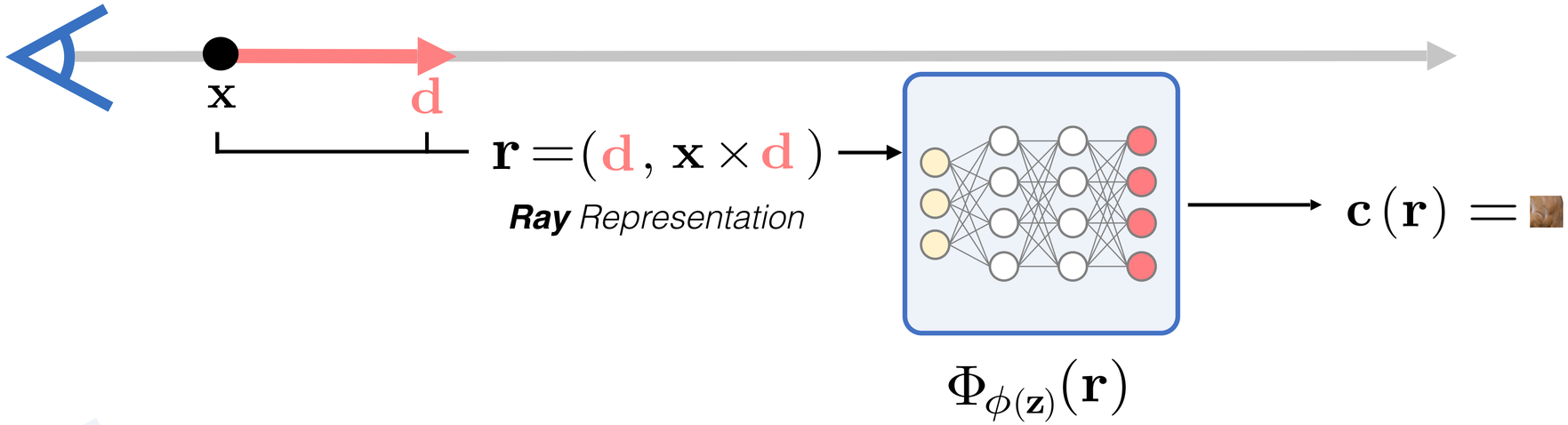
Cameron Smith
Visiting Researcher

Amortized (=feedforward, generalizable) 3D Reconstruction



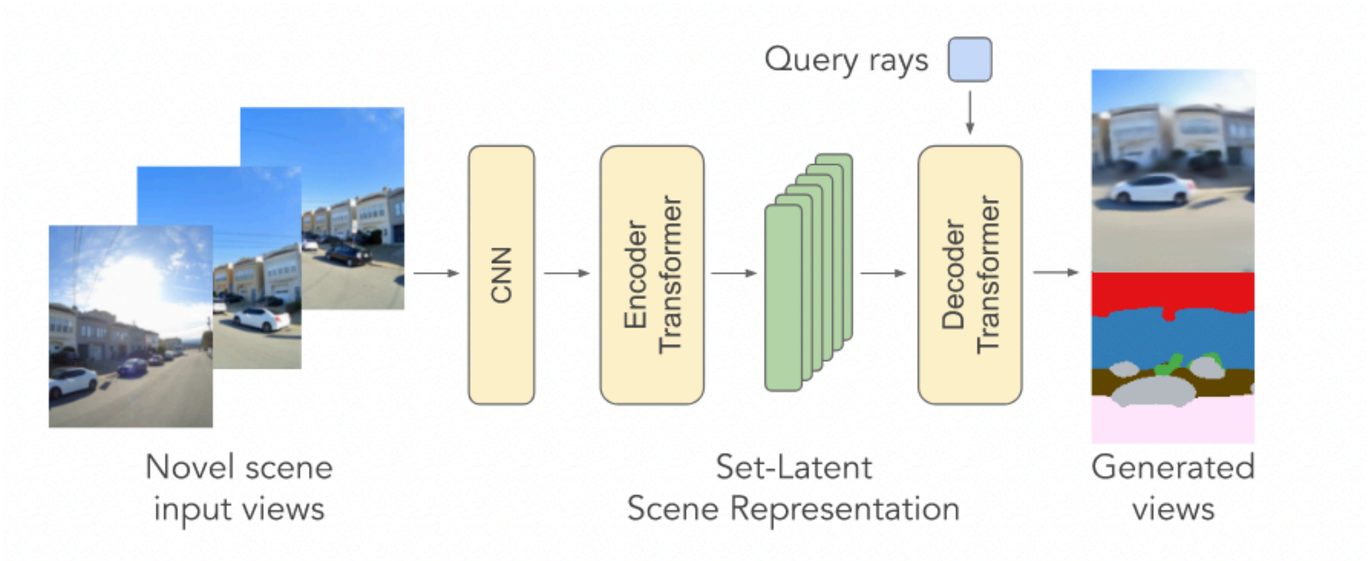
Light Field Networks





An Alternative Scene Representation

Light Fields Networks with Transformers: Scene Representation Transformer (CVPR 2022)

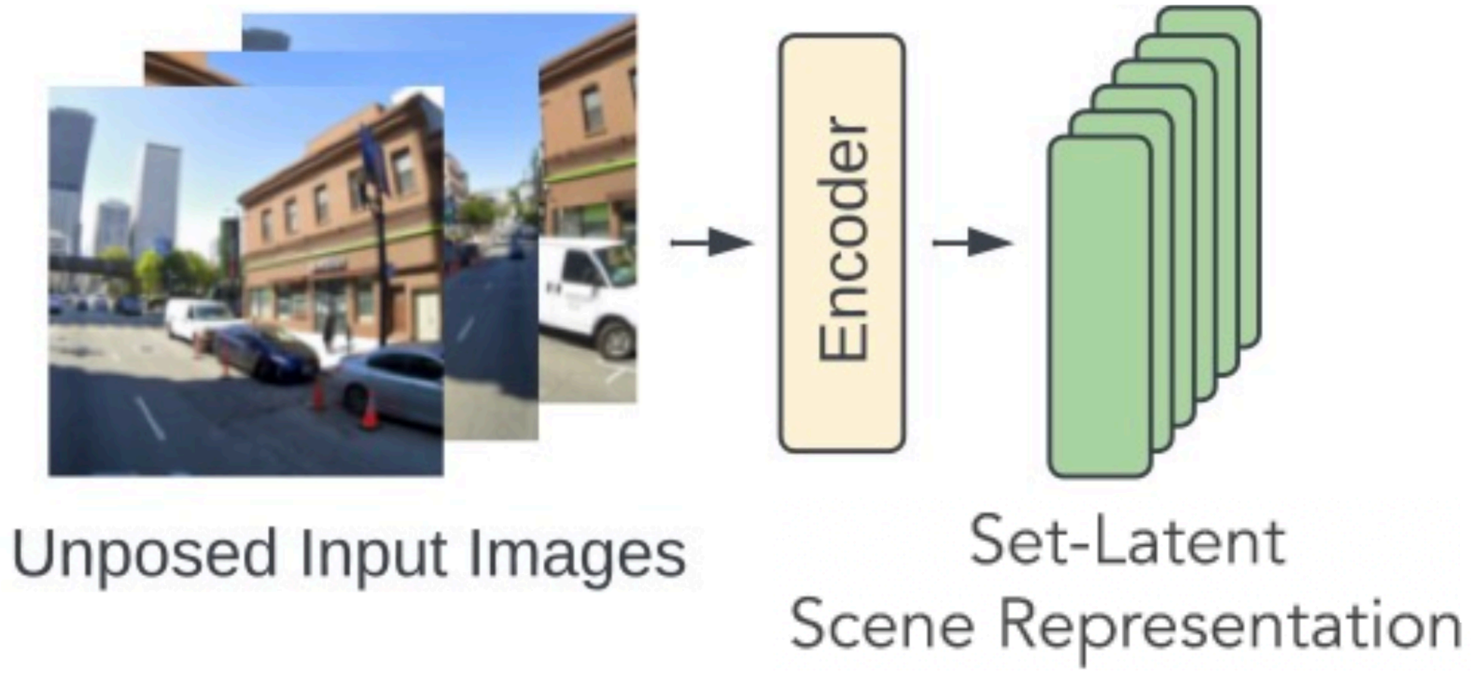


No 3D Renderer: Directly parameterizes Light Field!

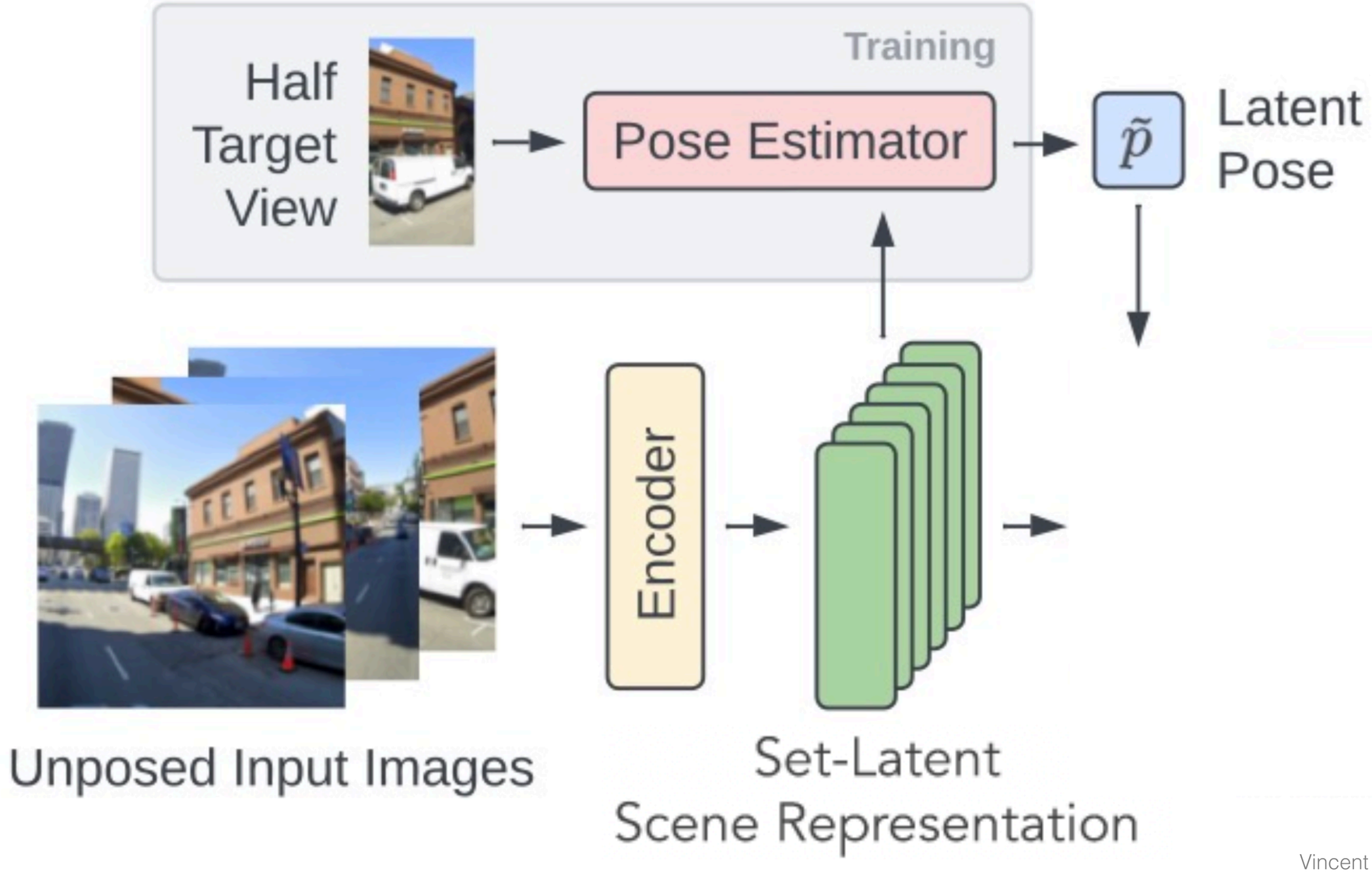
Scene Representation Transformer:
Geometry-Free Novel View Synthesis Through Set-Latent Scene Representations

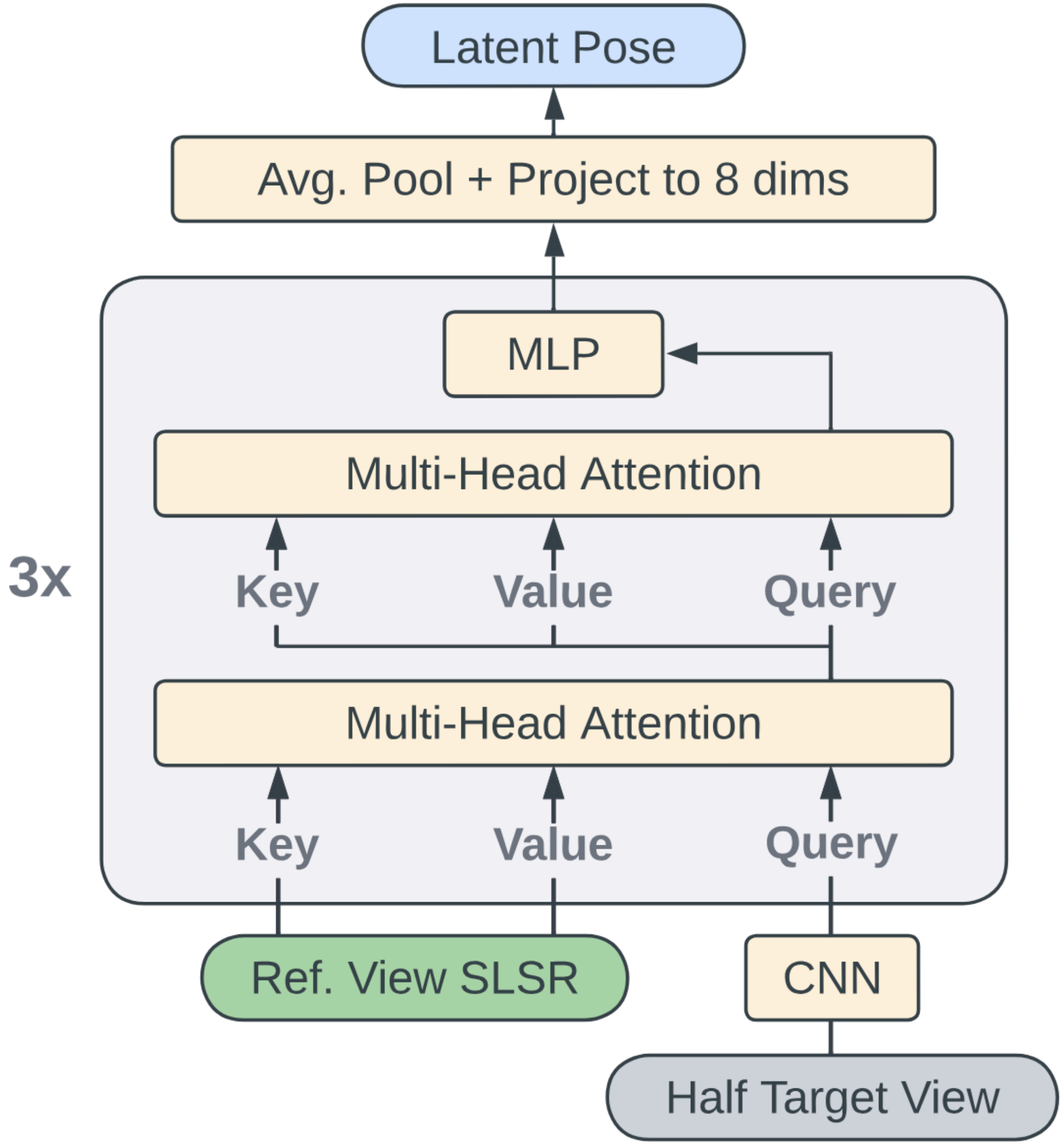
Mehdi S. M. Sajjadi Henning Meyer Etienne Pot Urs Bergmann Klaus Greff
Noha Radwan Suhani Vora Mario Lučić Daniel Duckworth Alexey Dosovitskiy*
Jakob Uszkoreit* Thomas Funkhouser Andrea Tagliasacchi‡*

RUST: Latent Neural Scene Representations from Unposed Imagery, Sajjadi et al. 2023

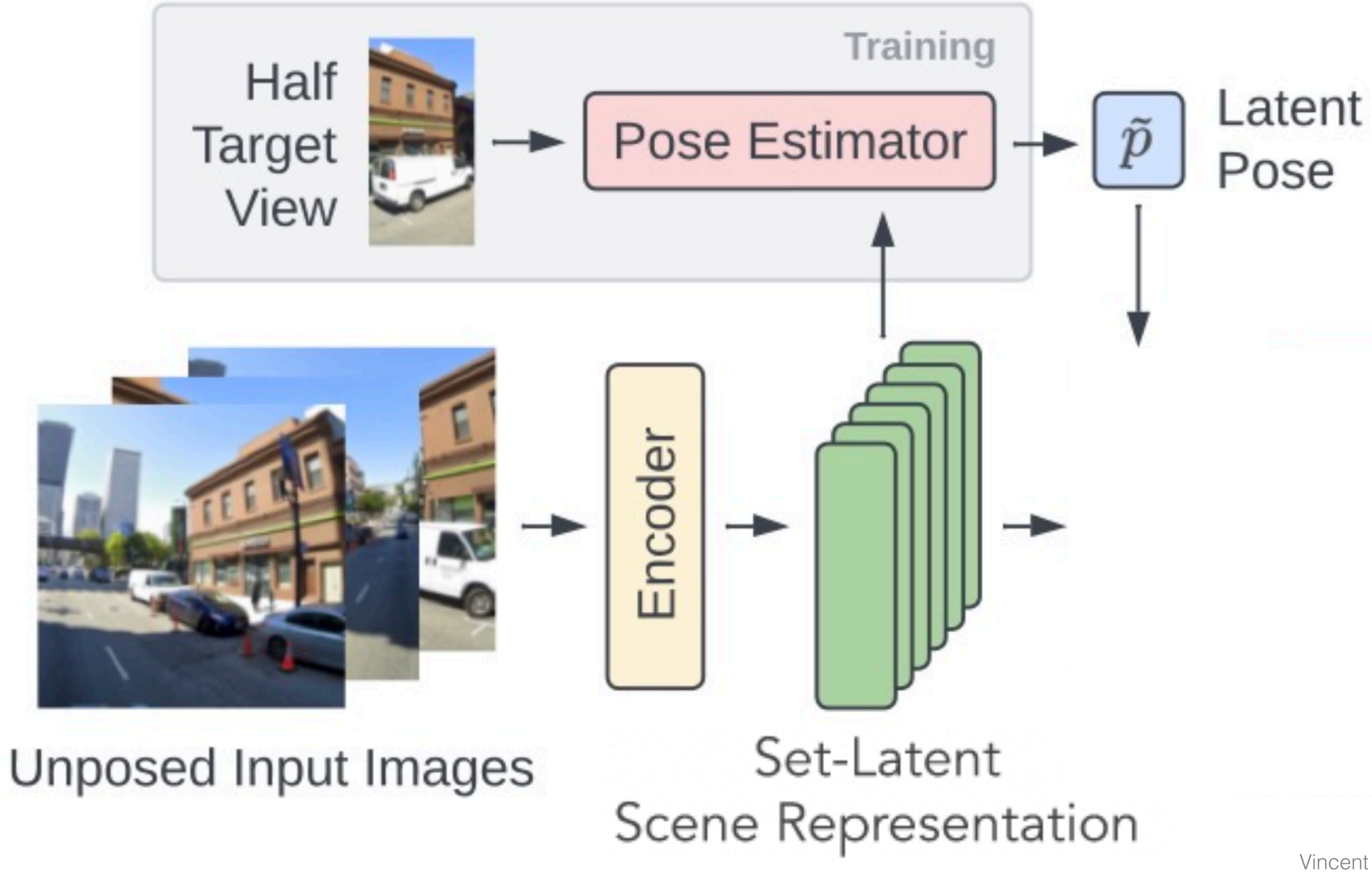


RUST: Latent Neural Scene Representations from Unposed Imagery, Sajjadi et al. 2023

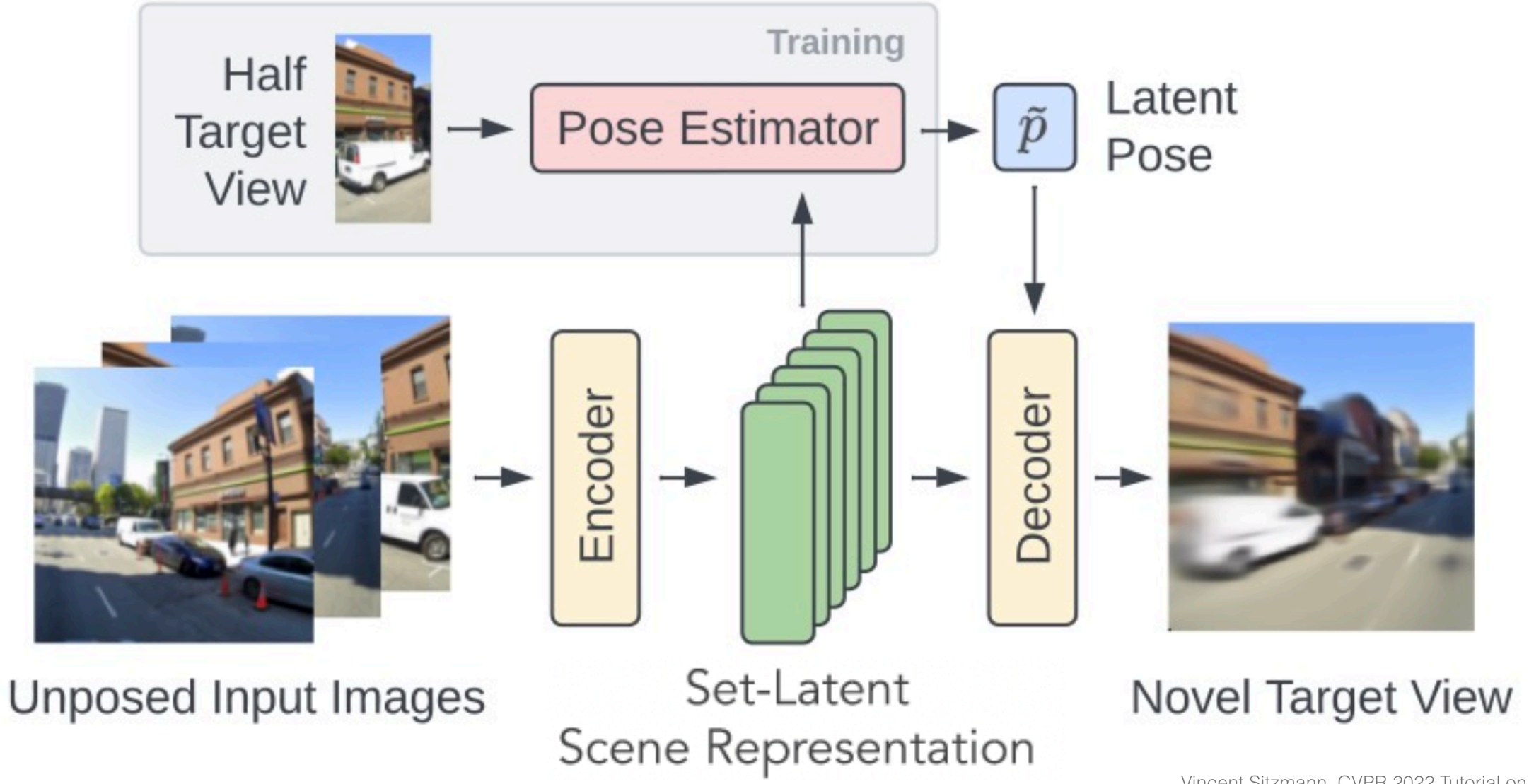



Latent Pose
Avg. Pool + Project to 8 dims
MLP
Multi-Head Attention
Key
Value
Query
3x
Multi-Head Attention
Key
Value
Query
Ref. View SLSR
CNN
Half Target View

RUST: Latent Neural Scene Representations from Unposed Imagery, Sajjadi et al. 2023

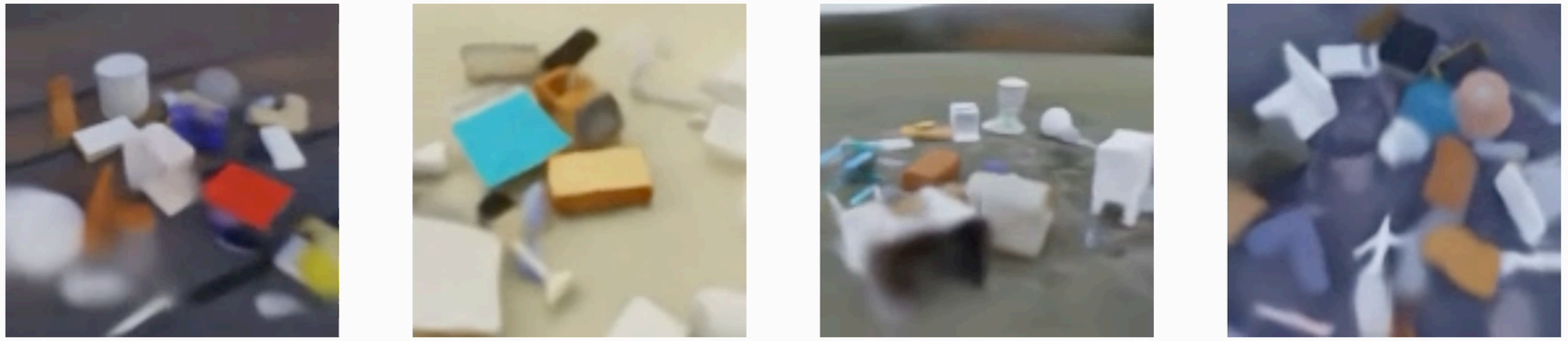


RUST: Latent Neural Scene Representations from Unposed Imagery, Sajjadi et al. 2023



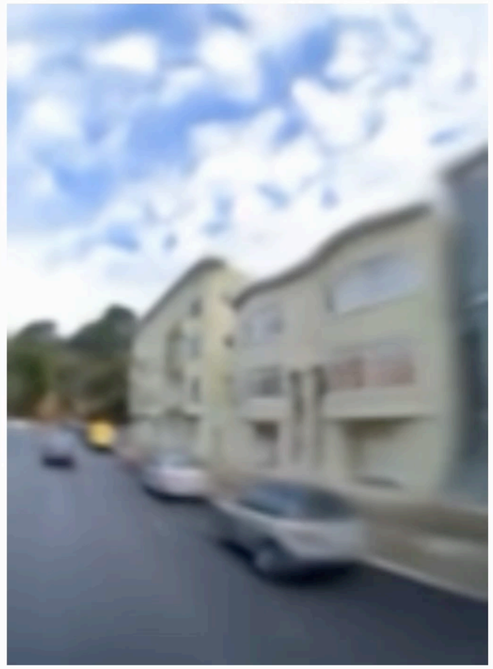
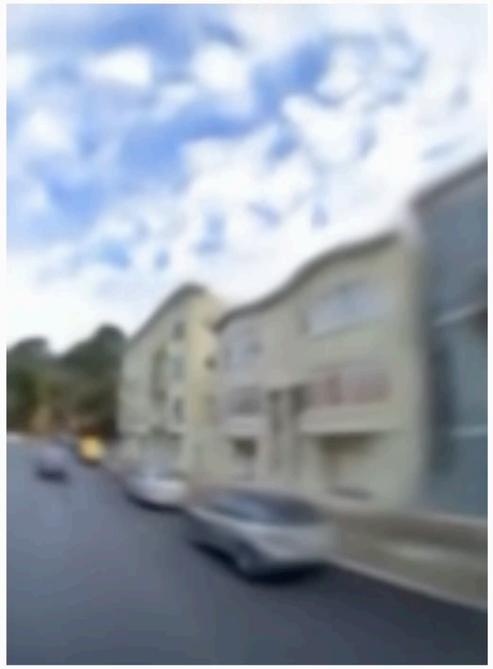
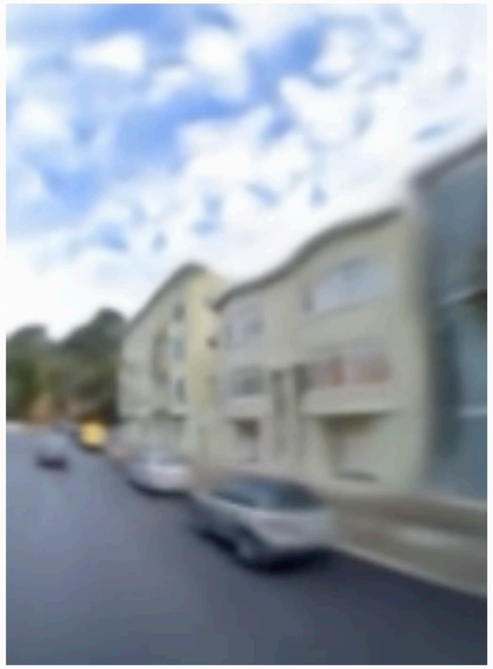
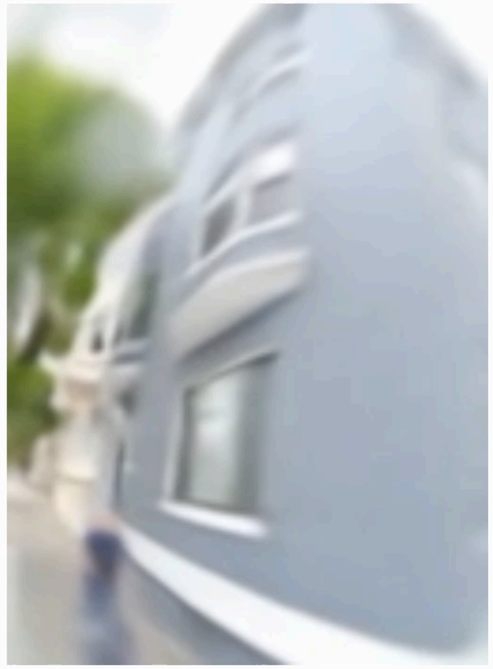
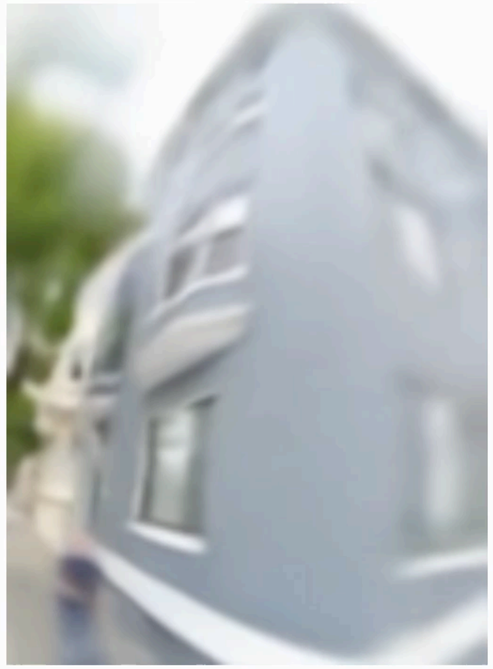
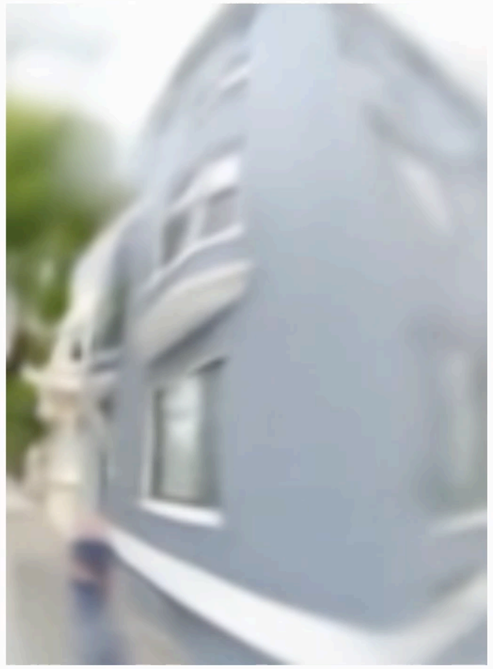
Camera movement induced by traversals in latent pose space. These are video equivalents of Fig. 5 (right) and Fig. 11 (right).

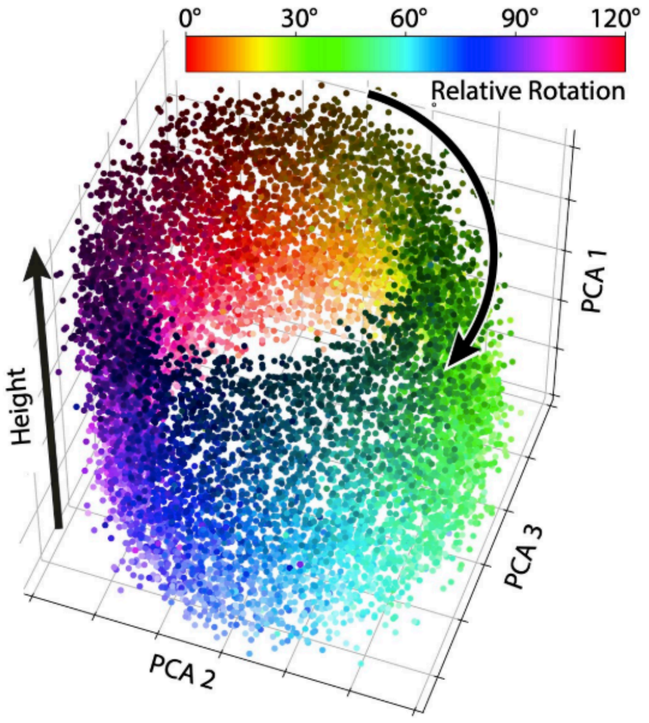
360 Rotation

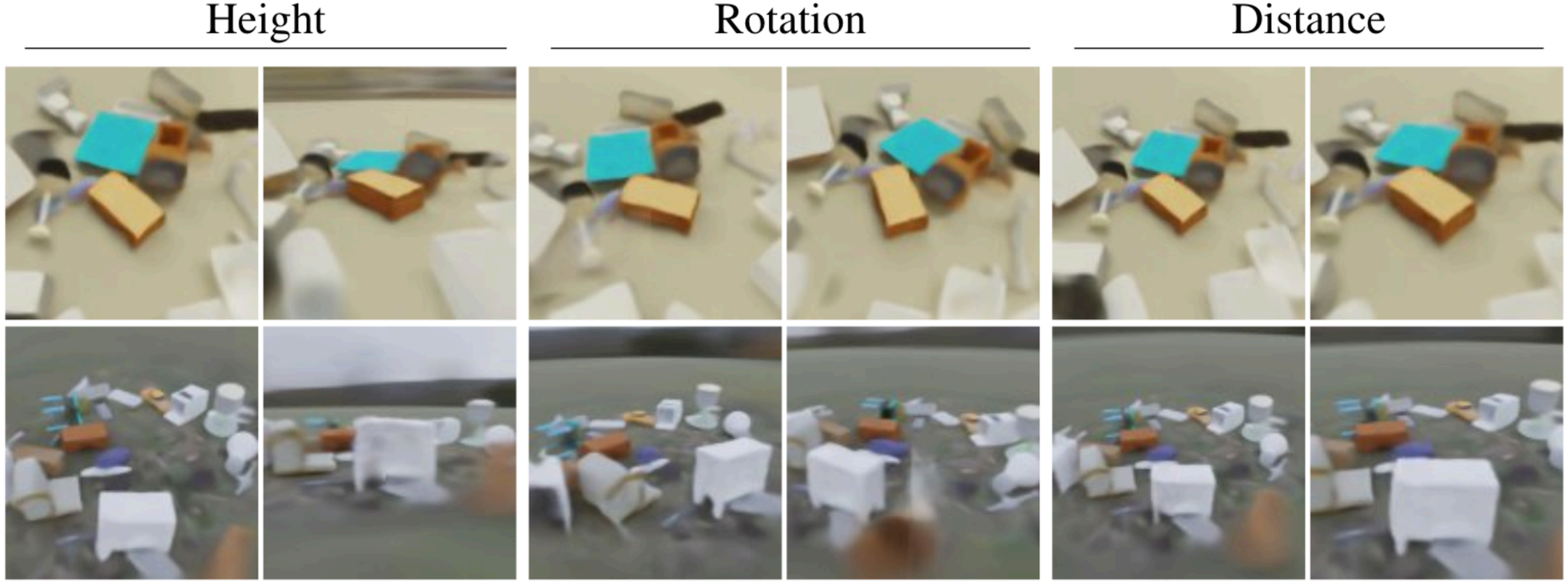


Distance

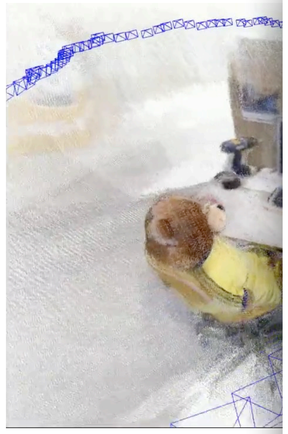
Height


0°
30°
60°
90°
120°
Relative Rotation
Height
PCA 1
PCA 3
PCA 2


Height
Rotation
Distance

Conventional Approaches



Prediction: This landscape will look different in ~2 years.

SFM & Sl **pred.**

- Per-sce
- If using
- SLAM: Prone to failure for particular trajectories, doesn't estimate camera intrinsics!