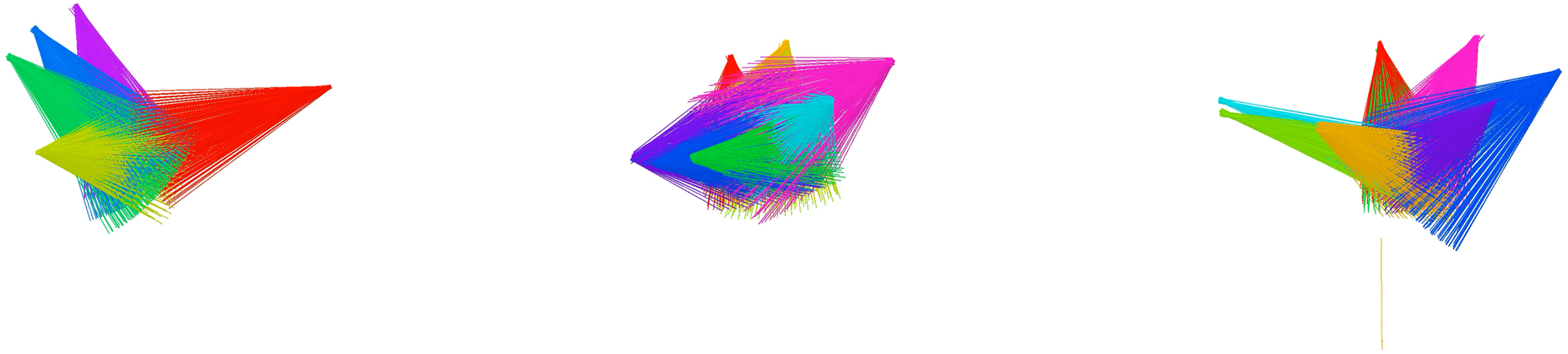
Rethinking Camera Parametrization for (Sparse-view) Pose Prediction

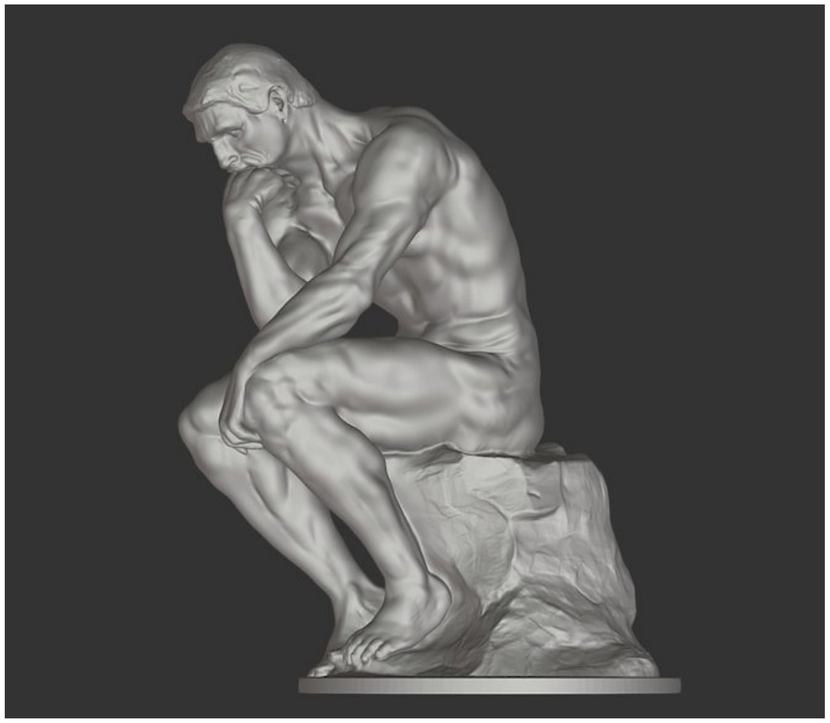


Shubham Tulsiani

Assistant Professor, CMU

Goal: Virtualize a real object

(for AR/VR, or posterity etc.)

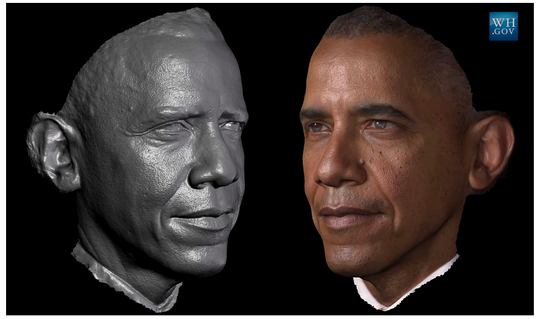


Dense Multi-view Images

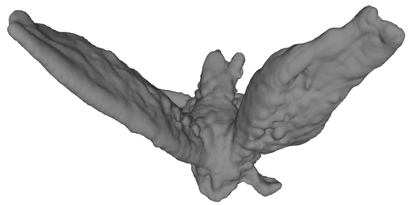
Single image ‘in-the-wild’

‘Ease’ of Capture

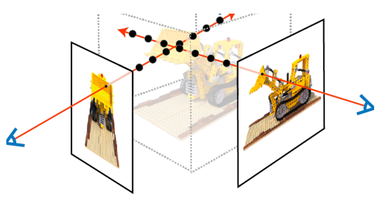
Faithfullness

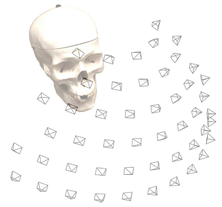


Precisely capture the underlying instance



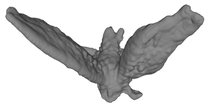
Hallucination/ Approximation

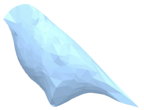




NeRF, IDR, NeUS,
Gaussian Splatting etc.

Sparse-view 3D in-the-wild





CMR, Zero-1-to-3,
TripoSR etc.

Faithfullness

'Ease' of Inference

(circa 2021)

Jason: *I want to reconstruct Craigslist objects*

me: *Ok, let's first setup a NeRF baseline. Run SfM to get poses, and …*

Jason (a couple days later): *Well, I tried really hard, but COLMAP keeps crashing…*

3D from images: the typical recipe

Images

COLMAP

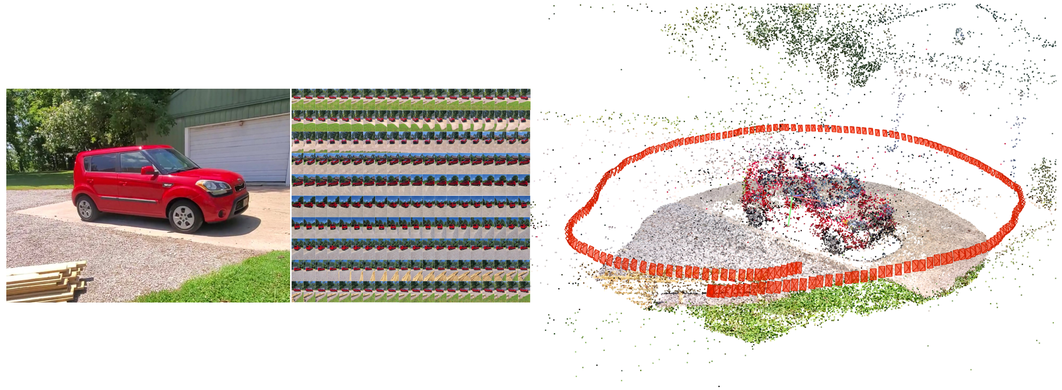
Cameras

Diff. Rendering + Optimization

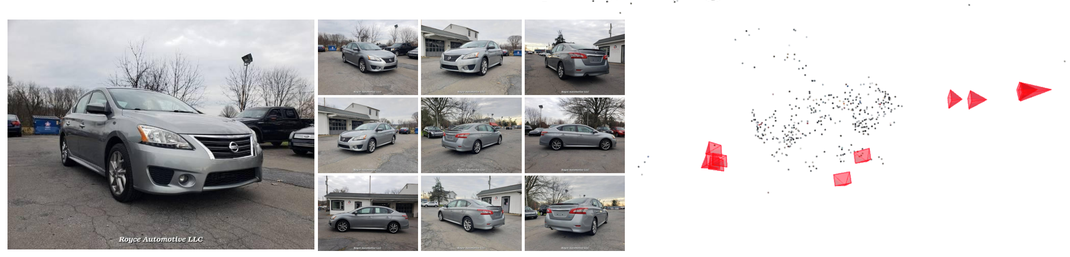
3D Representation

Challenge: Classical SfM is not robust for Sparse-view Pose

202 Input Images



10 Input Images



Images

COLMAP

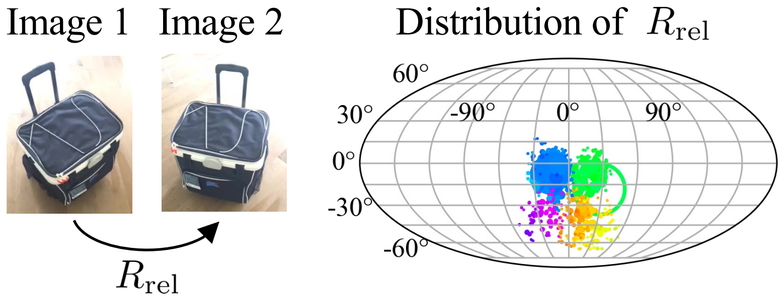
Cameras

Diff. Rendering + Optimization

3D Representation

Let's learn sparse-view pose prediction!

*essentially, learning-based localization (without mapping)

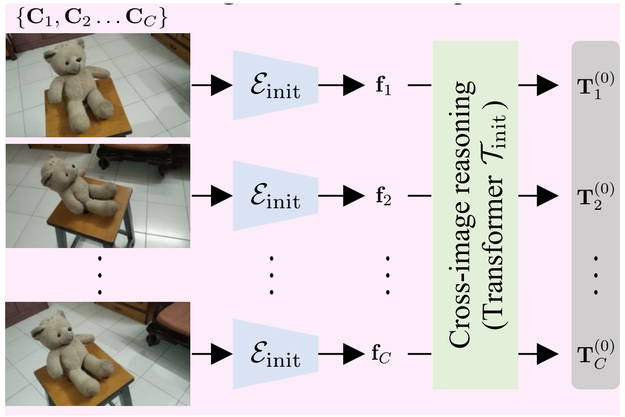


RelPose, RelPose++ *(Lin*, Zhang* et. al.)*

$\{\mathbf{I}_n\}$

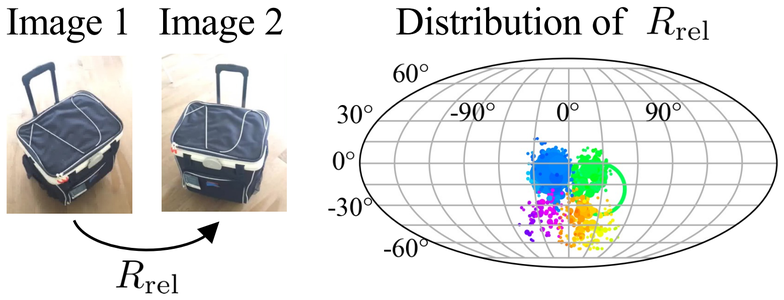
$\{(R_n, \mathbf{t}_n)\}$

Let's learn sparse-view pose prediction!



SparsePose *(Sinha et. al.)*

Regression + 3D-based refinement



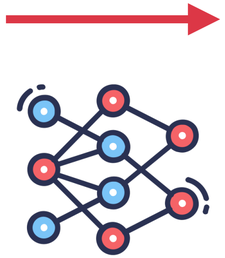
RelPose, RelPose++ *(Lin*, Zhang* et. al.)*

Energy-based modeling

PoseDiffusion *(Wang et. al.)*

Denoising diffusion

$\{\mathbf{I}_n\}$



$\{(R_n, \mathbf{t}_n)\}$

Innovations in uncertainty modeling, joint 3D inference, etc..

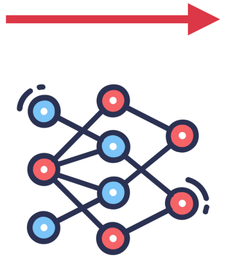
Parametrization for Prediction: global extrinsics (+ intrinsics)

But is this a good representation for learning-based prediction?

(difficult for global models to leverage local cues e.g. correspondences)

Can we re-parametrize camera estimation as a local prediction task?

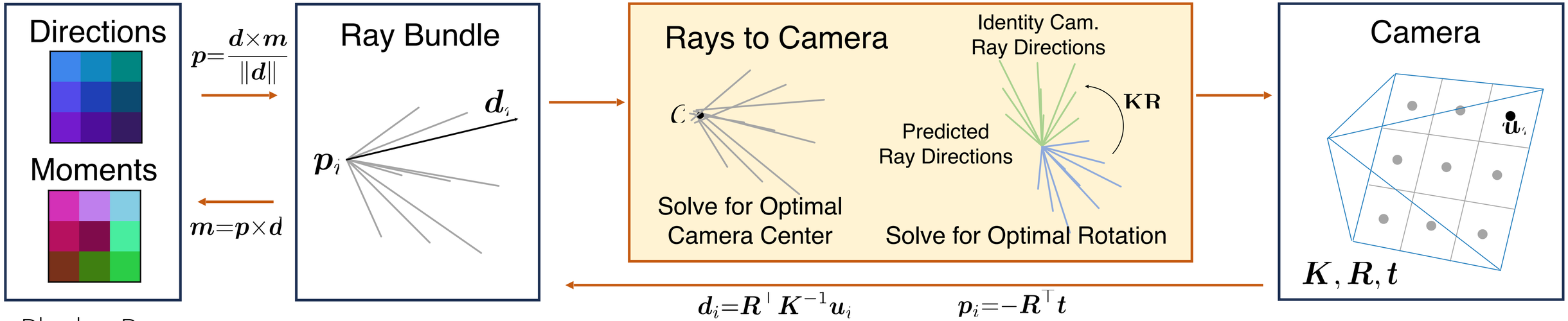
$\{\mathbf{I}_n\}$



$\{(R_n, \mathbf{t}_n)\}$

Representing Cameras via Ray Bundles

Key Idea: Represent cameras via per-pixel rays in a common coordinate frame



Plucker Ray Parametrization

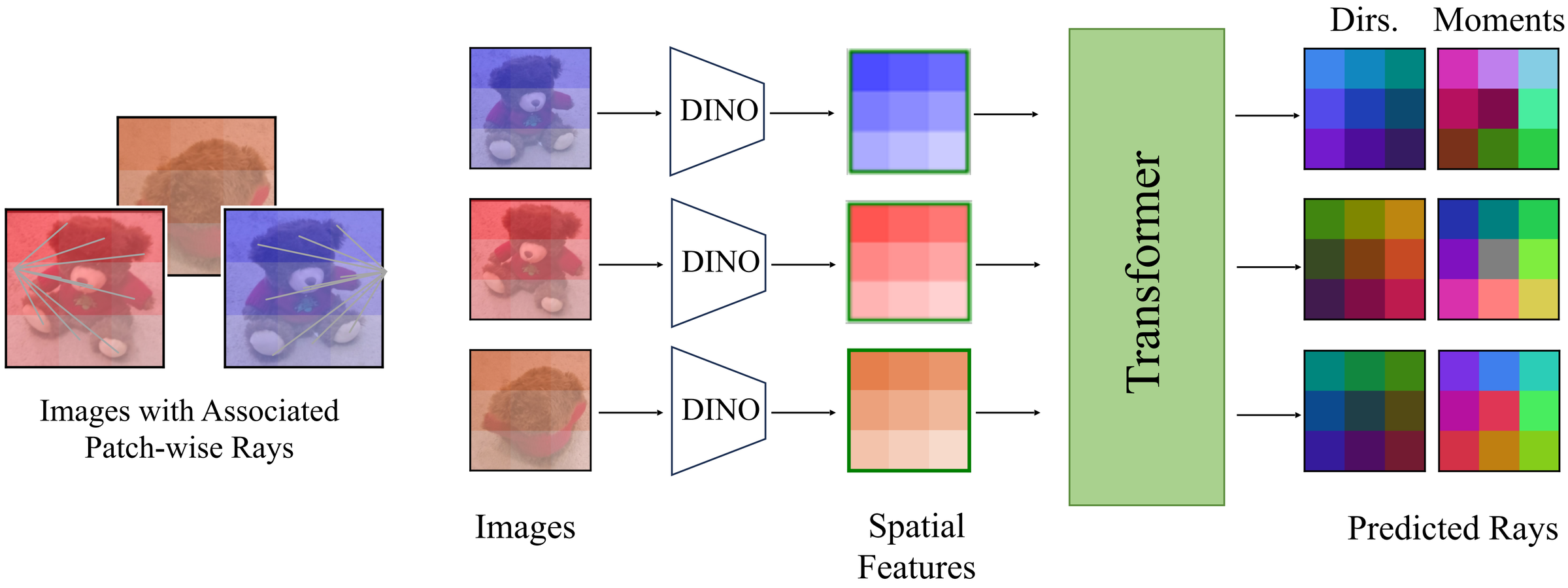
A **distributed** and **generic** representation

(inspired by work in camera calibration e.g. Grossberg & Nayar, Schops et. al.)

Can analytically recover pinhole camera parameters given predicted rays

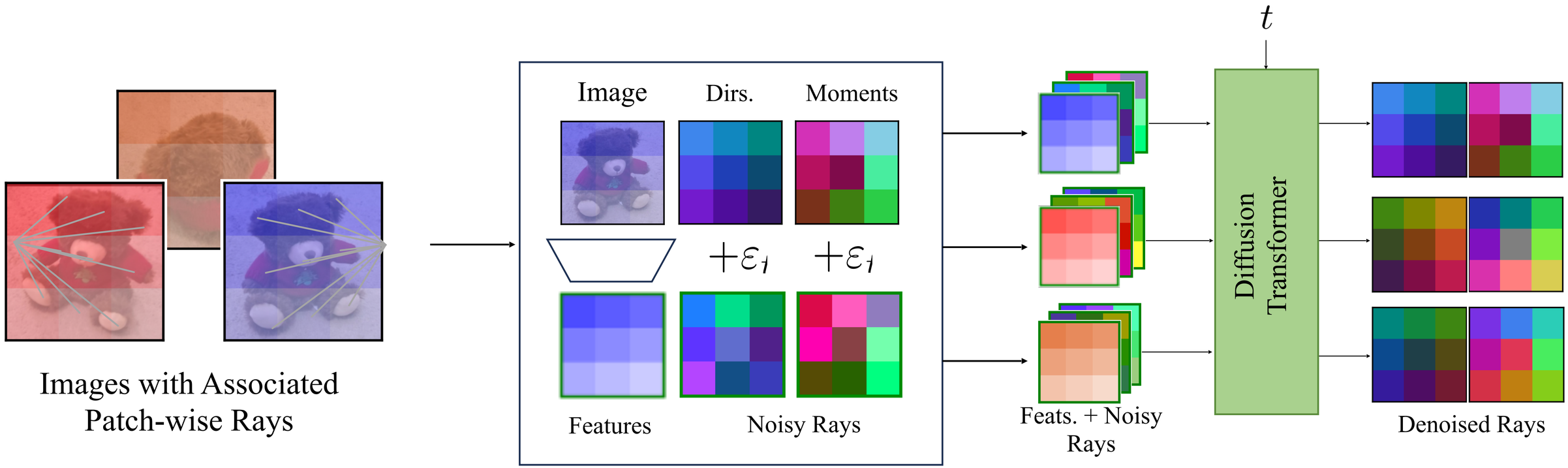
Cameras as Rays: Pose Estimation via Ray Diffusion. *Zhang*, Lin*, Ramanan, Tulsiani.* In ICLR 2024

Camera Prediction via Ray Regression



Cameras as Rays: Pose Estimation via Ray Diffusion. *Zhang*, Lin*, Ramanan, Tulsiani.* In ICLR 2024

Camera Prediction via Ray Diffusion

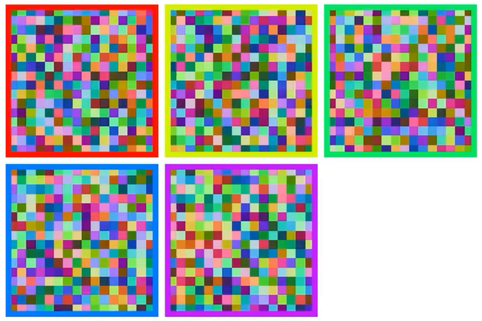


Cameras as Rays: Pose Estimation via Ray Diffusion. *Zhang*, Lin*, Ramanan, Tulsiani.* In ICLR 2024

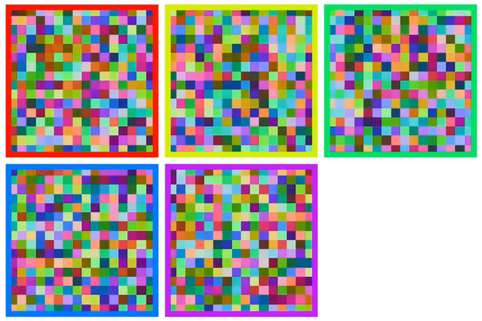
Visualizing Reverse Diffusion

Input Images

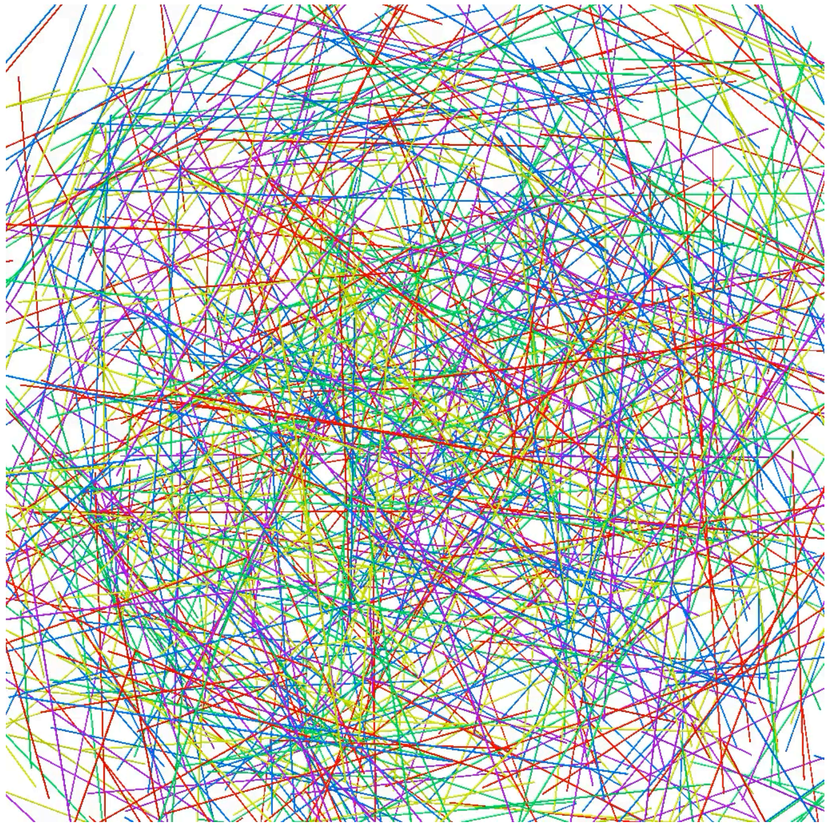
Directions



Moments



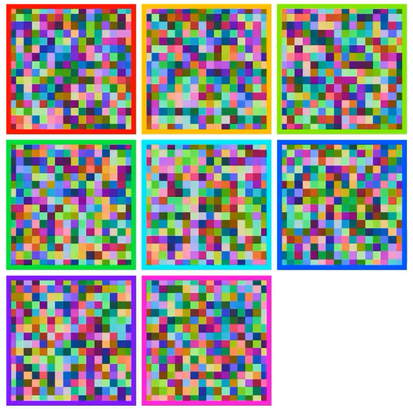
3D Rays



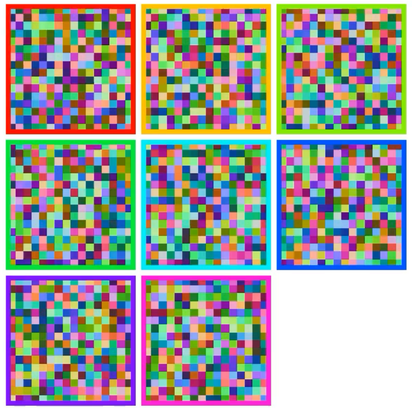
Visualizing Reverse Diffusion

Input Images

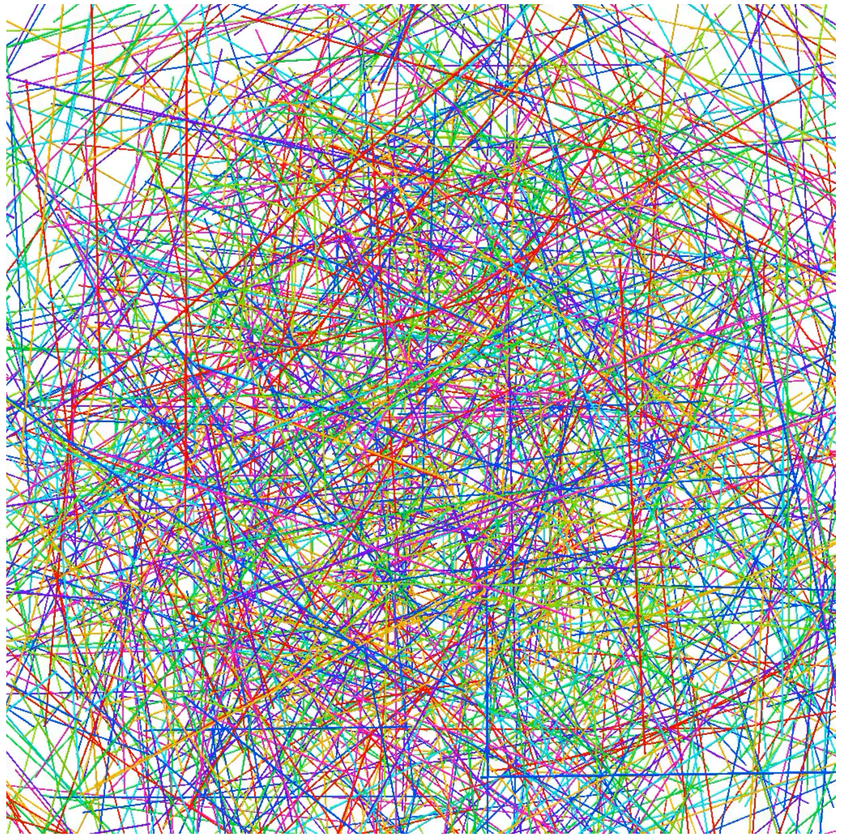
Directions



Moments



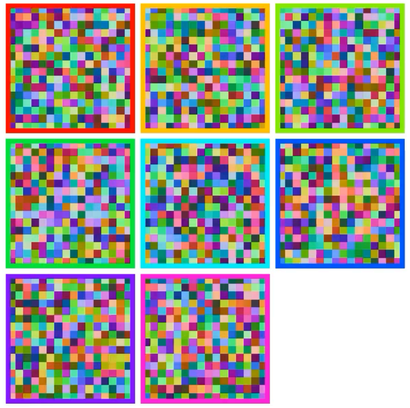
3D Rays



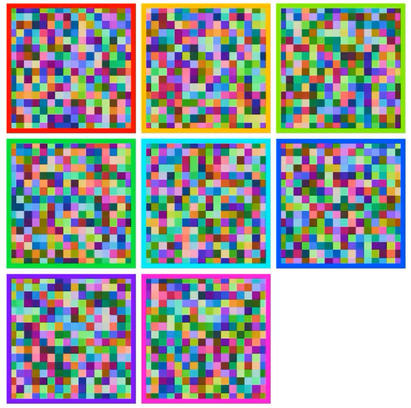
Visualizing Reverse Diffusion

Input Images

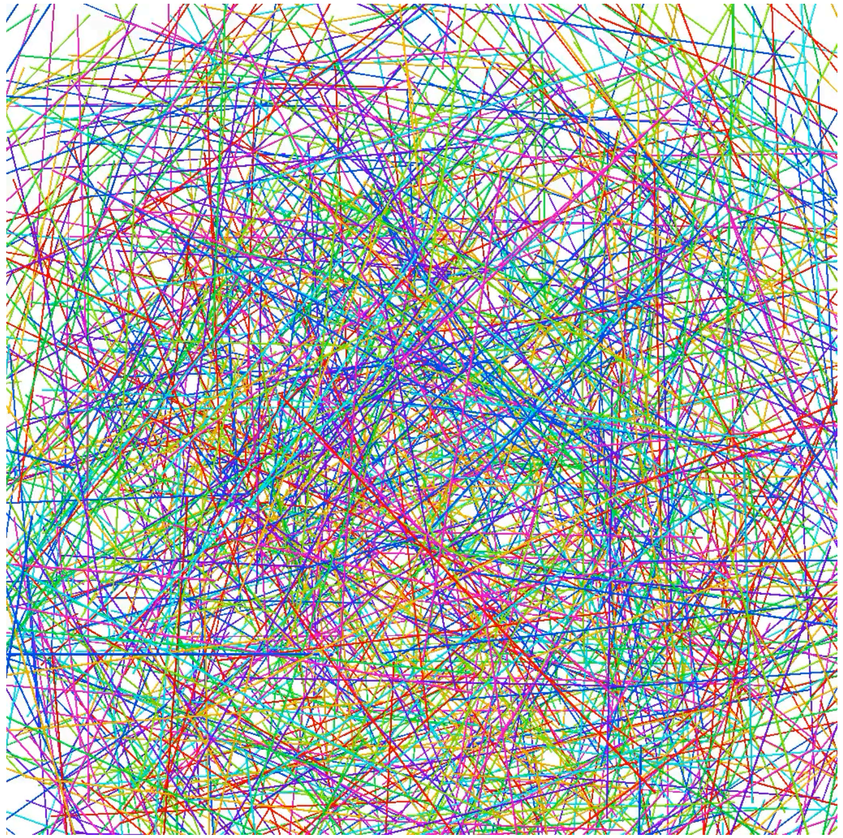
Directions

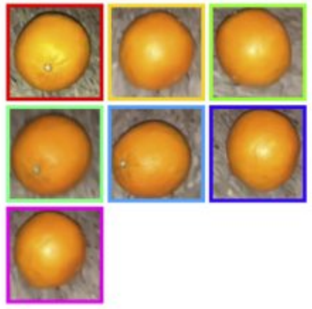
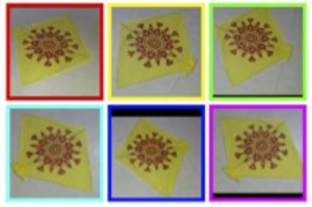


Moments

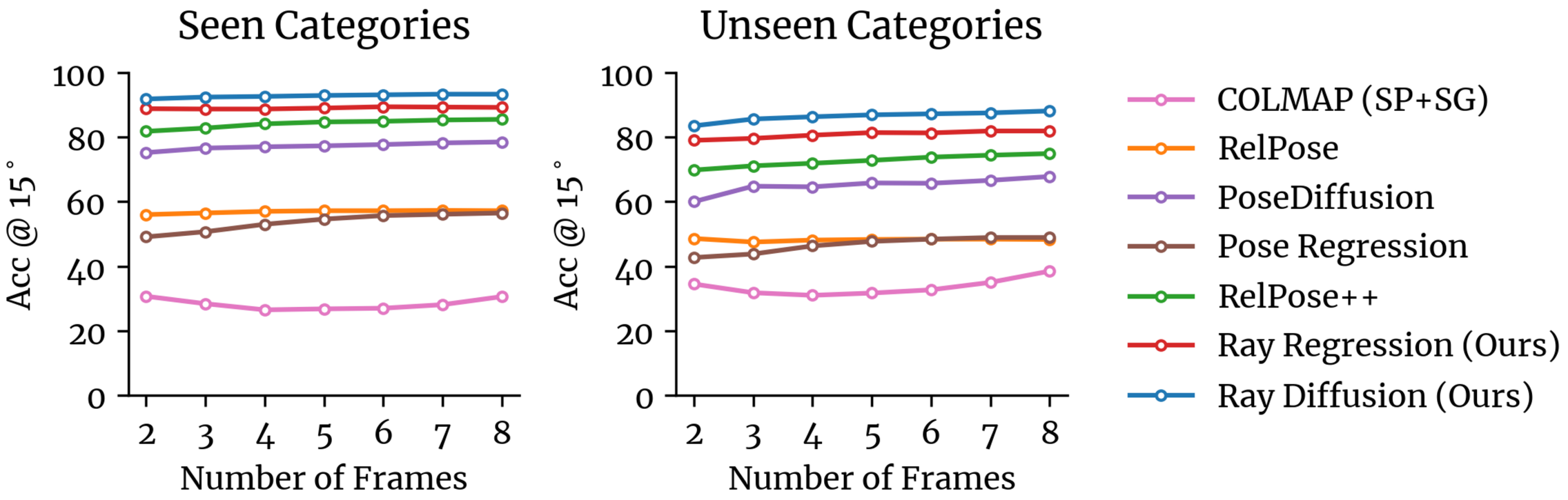


3D Rays

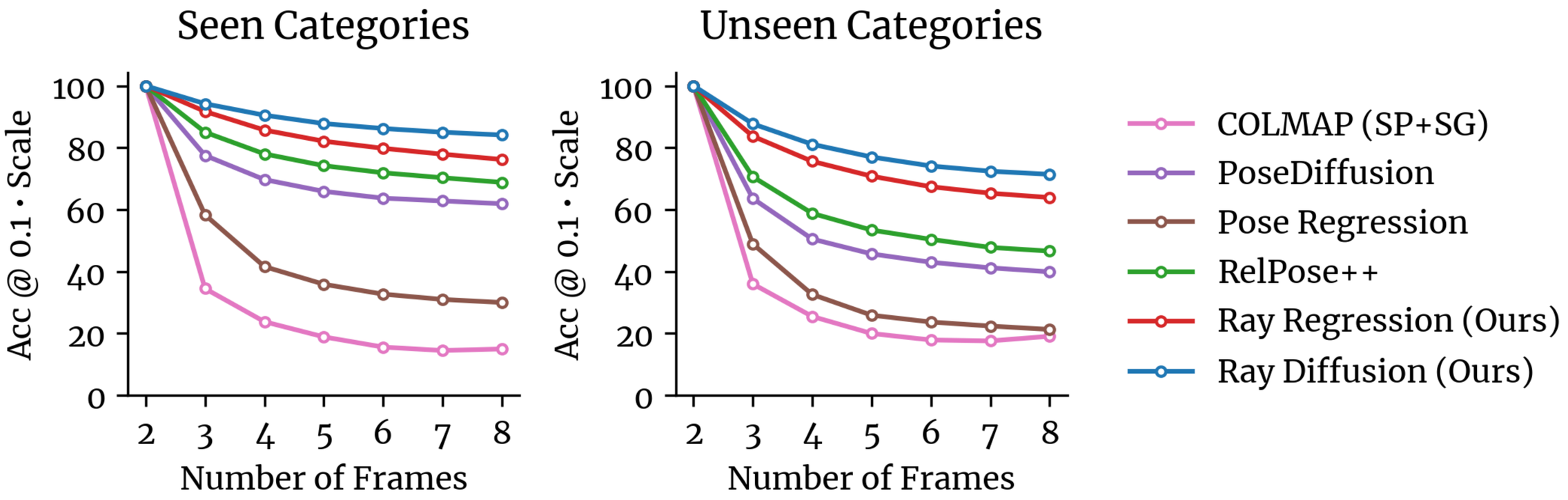


Images	Pose Diffusion	RelPose++	Ray Regression (Ours)	Ray Diffusion (Ours)
				
				
				
				

Quantitative: Rotation Accuracy (% < 15°)

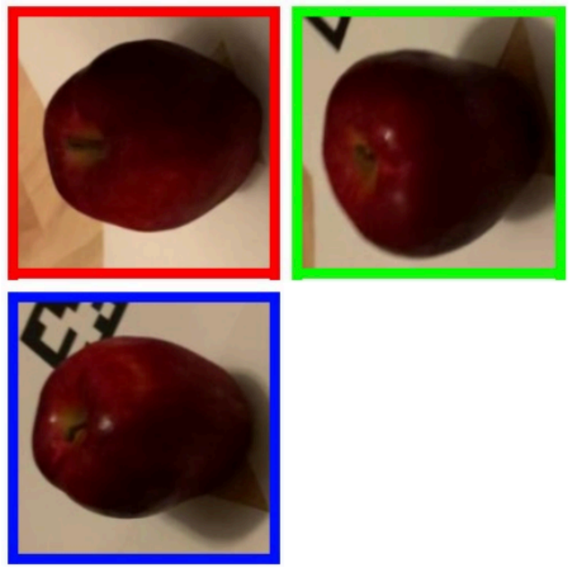


Quantitative: Cam Center Accuracy (% < 0.1)

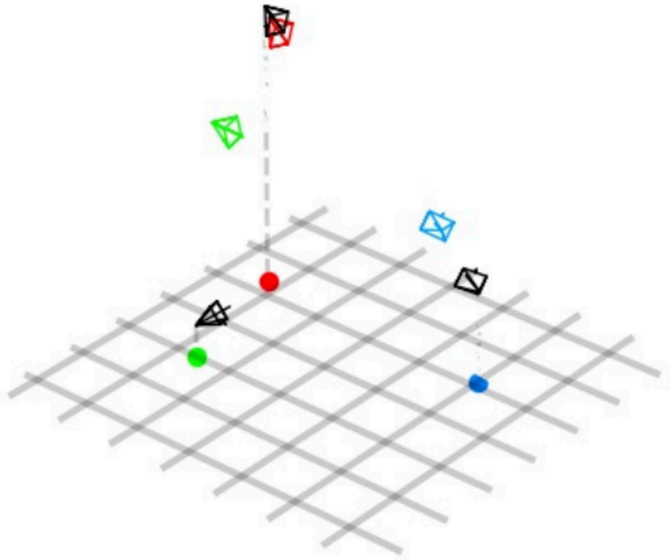


Modeling Uncertainty

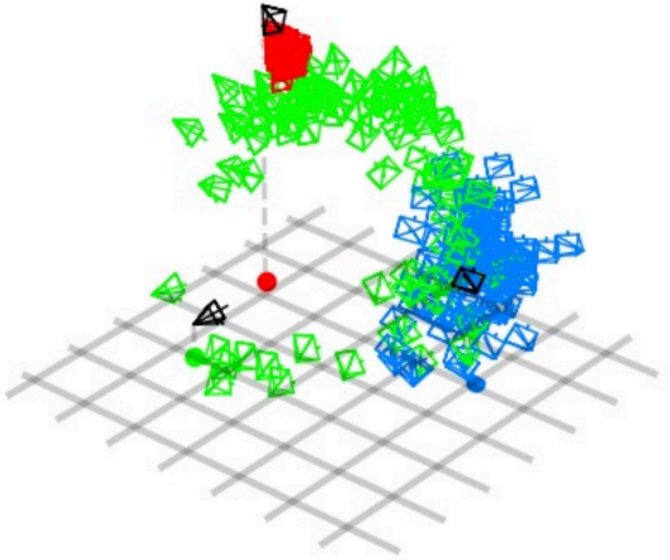
Input Images



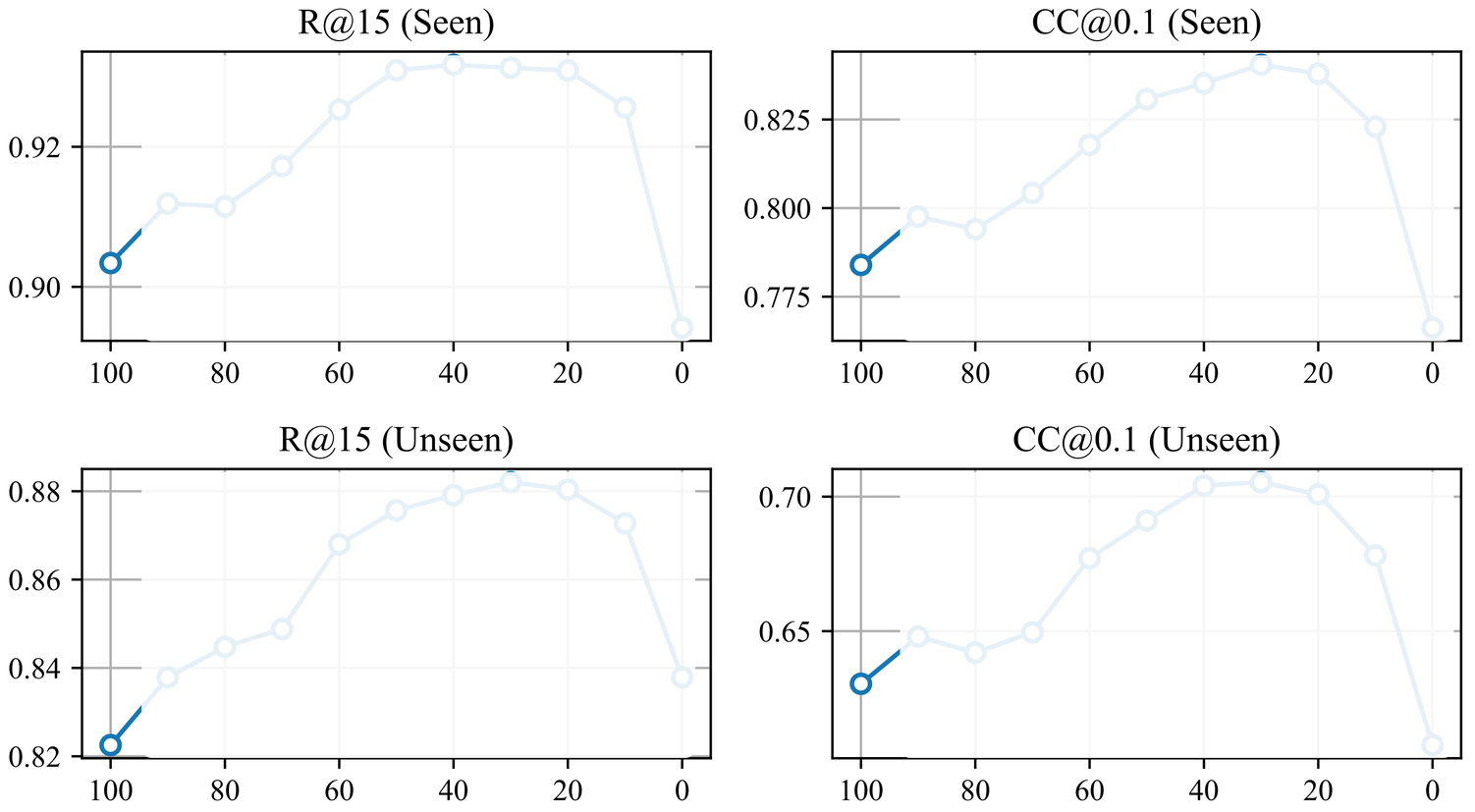
Ray Regression



Ray Diffusion (100 Samples)



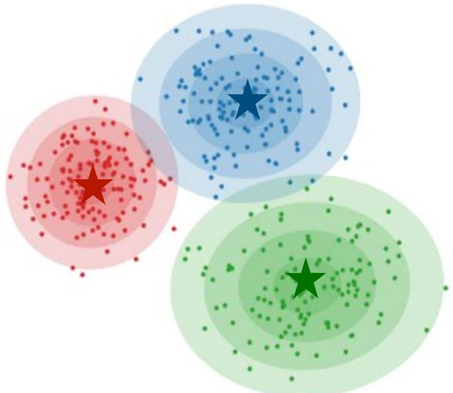
Early Stopping in Reverse Diffusion



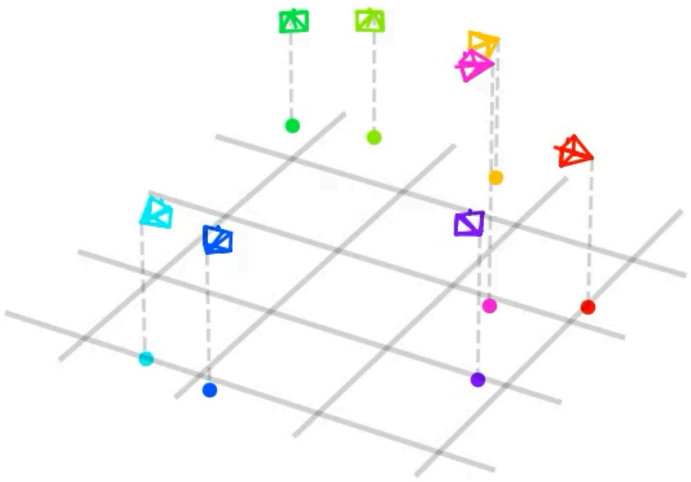
At each diffusion timestep, we obtain a denoised ray prediction

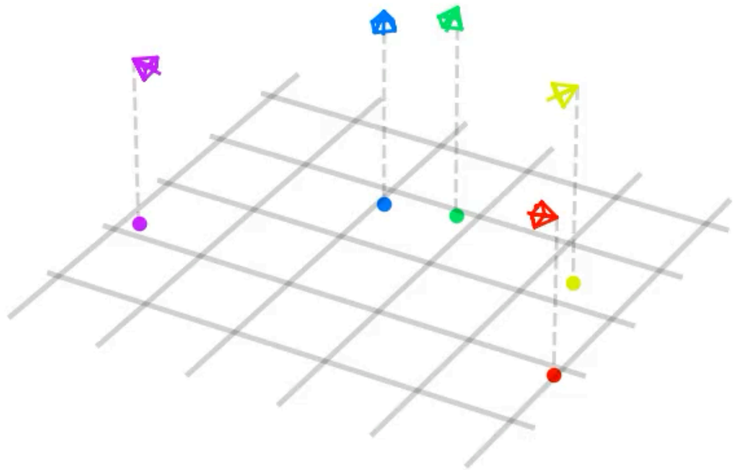
Pose accuracy increases initially, but drastically drops later

***Hypothesis**: For accuracy metrics, we want modes, not samples*

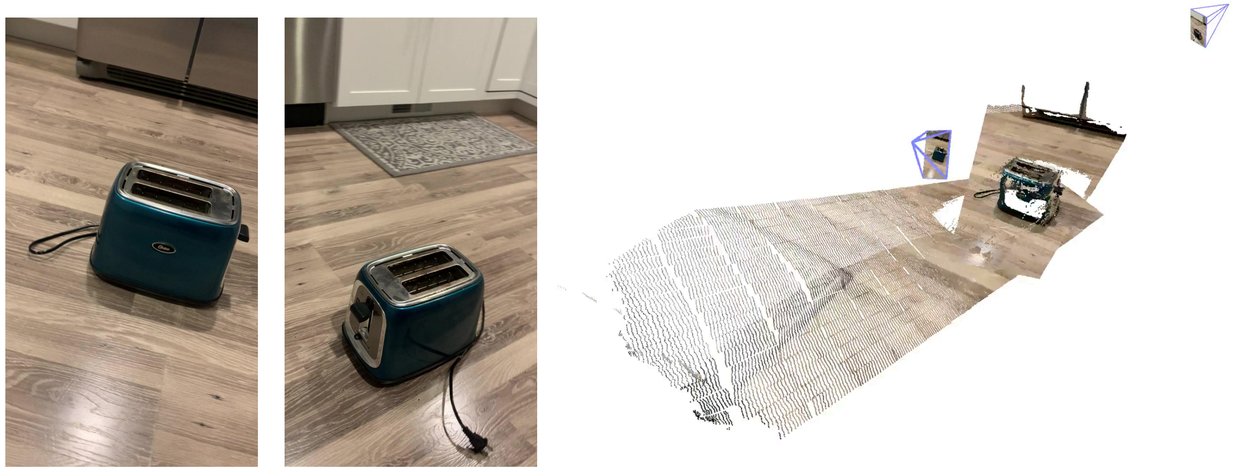


In-the-wild Generalization



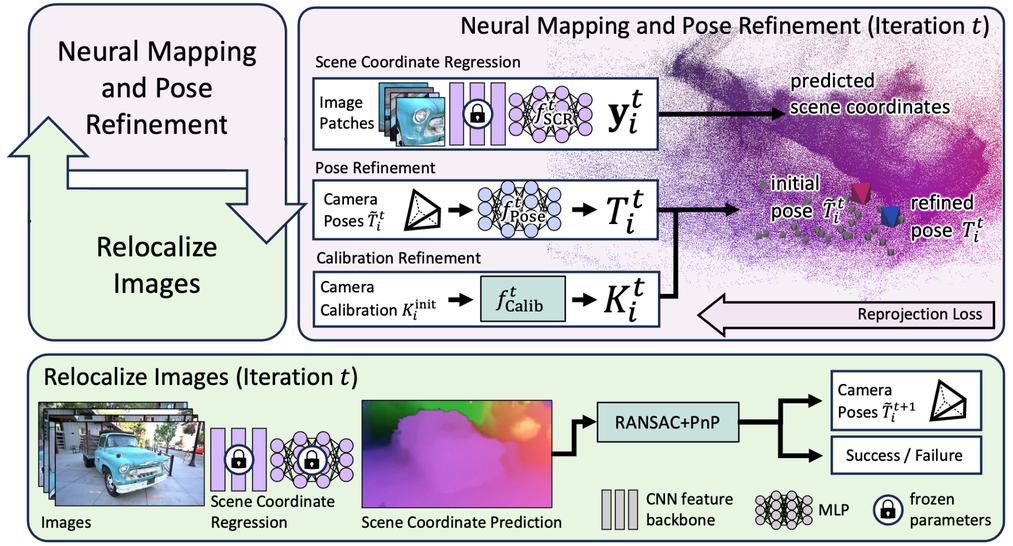


Dense Prediction for Pose (and 3D) Estimation



Dust3r (CVPR 24), Mast3r (ECCV 24)

Pointmap prediction from image pairs + global alignment



Ace-Zero (ECCV 24)

Rethinking SfM pipeline via Scene coordinate regression

Were cameras the only missing piece for Sparse-view 3D?

Images

RayDiffusion, Dust3r

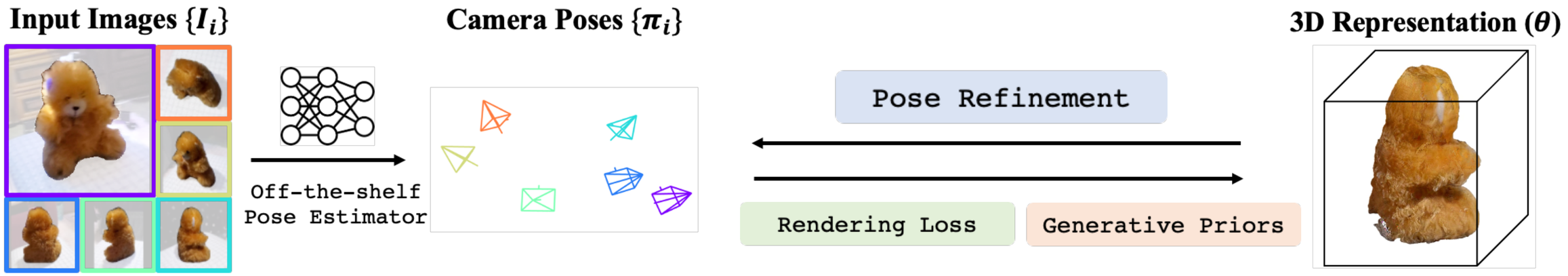
Cameras

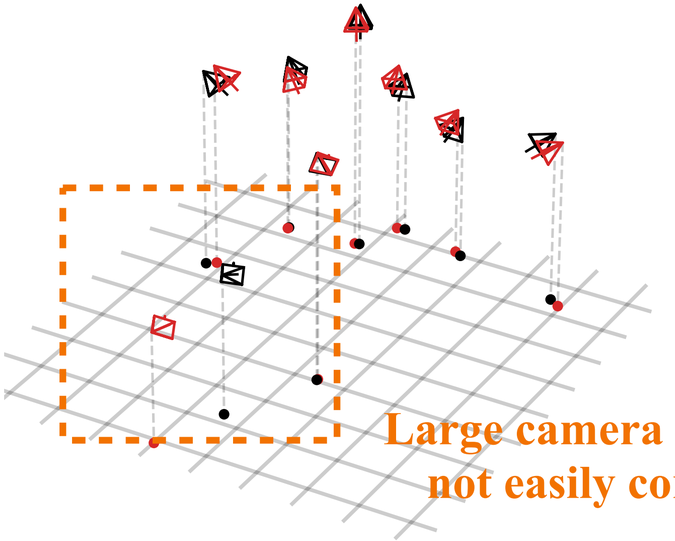
Diff. Rendering + Optimization

3D Representation

Challenges due to unobserved aspects, pose outliers

Analysis by Generative Synthesis

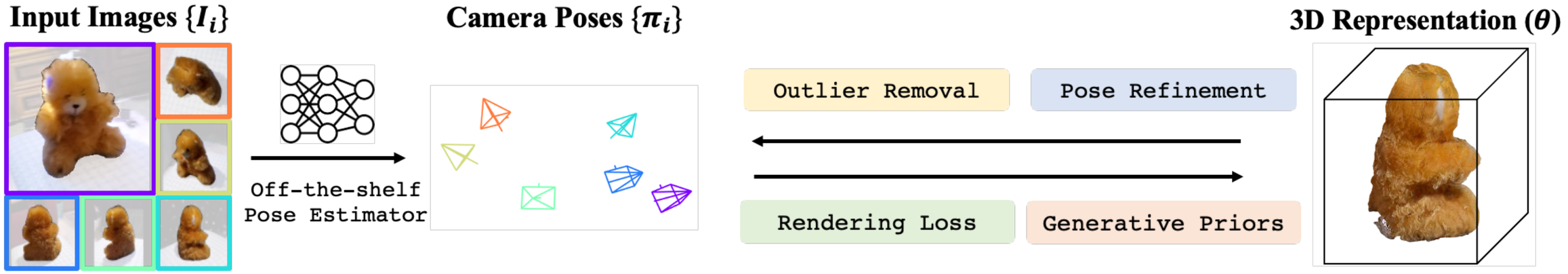


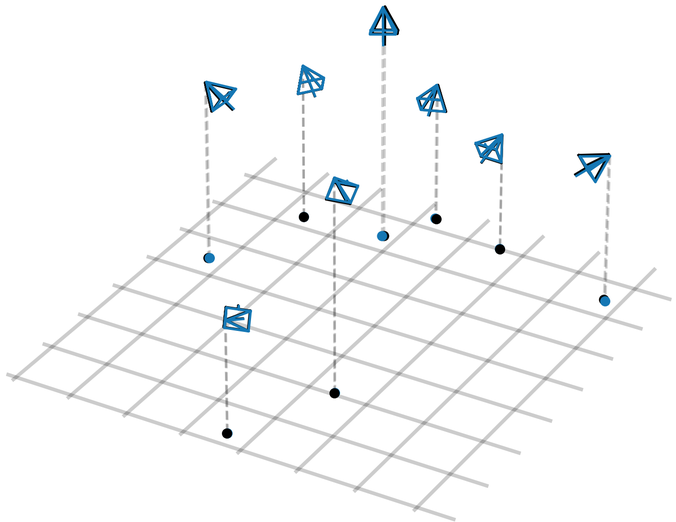


Large camera errors are not easily corrected!

Sparse-view Pose Estimation and Reconstruction via Analysis by Generative Synthesis. *Zhao & Tulsiani, In NeurIPS 2024 (on Arxiv soon)*

Analysis by Generative Synthesis





Sparse-view Pose Estimation and Reconstruction via Analysis by Generative Synthesis. *Zhao & Tulsiani, In NeurIPS 2024 (on Arxiv soon)*

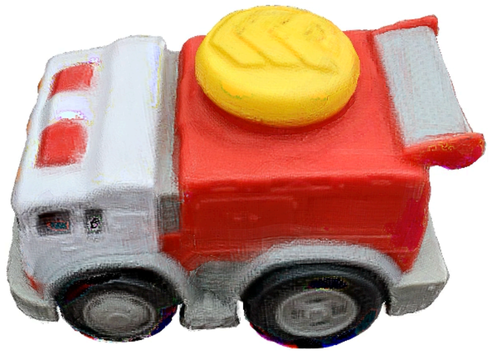
Piecing it together: Analysis by Generative Synthesis

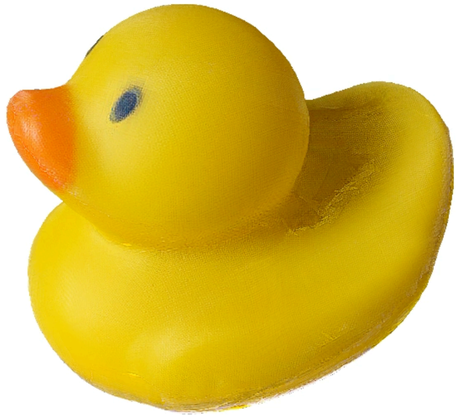
Sparse-view Pose Estimation and Reconstruction via Analysis by Generative Synthesis. *Zhao & Tulsiani, In NeurIPS 2024 (on Arxiv soon)*

Piecing it together: Analysis by Generative Synthesis

Sparse-view Pose Estimation and Reconstruction via Analysis by Generative Synthesis. *Zhao & Tulsiani, In NeurIPS 2024 (on Arxiv soon)*

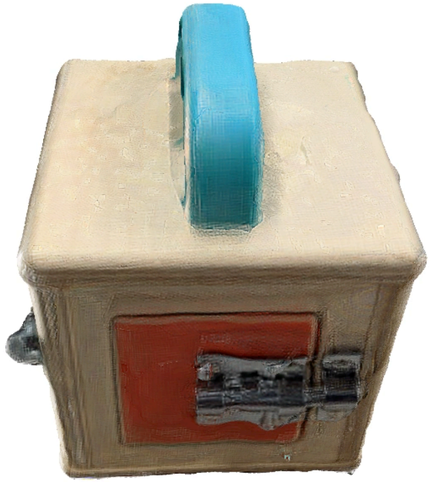
Piecing it together: Analysis by Generative Synthesis





Sparse-view Pose Estimation and Reconstruction via Analysis by Generative Synthesis. *Zhao & Tulsiani, In NeurIPS 2024 (on Arxiv soon)*

Piecing it together: Analysis by Generative Synthesis



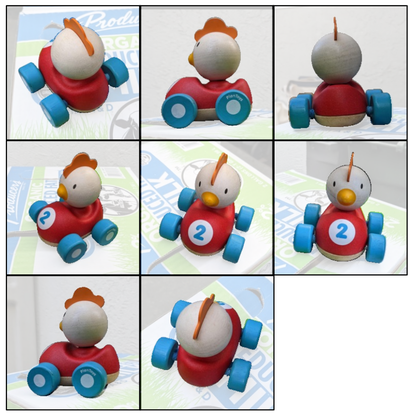
Sparse-view Pose Estimation and Reconstruction via Analysis by Generative Synthesis. *Zhao & Tulsiani, In NeurIPS 2024 (on Arxiv soon)*

Input

Initial Pose

SPARF

Ours (w/ 3D)



Thank you!

Jason

Amy

Qitao

Deva