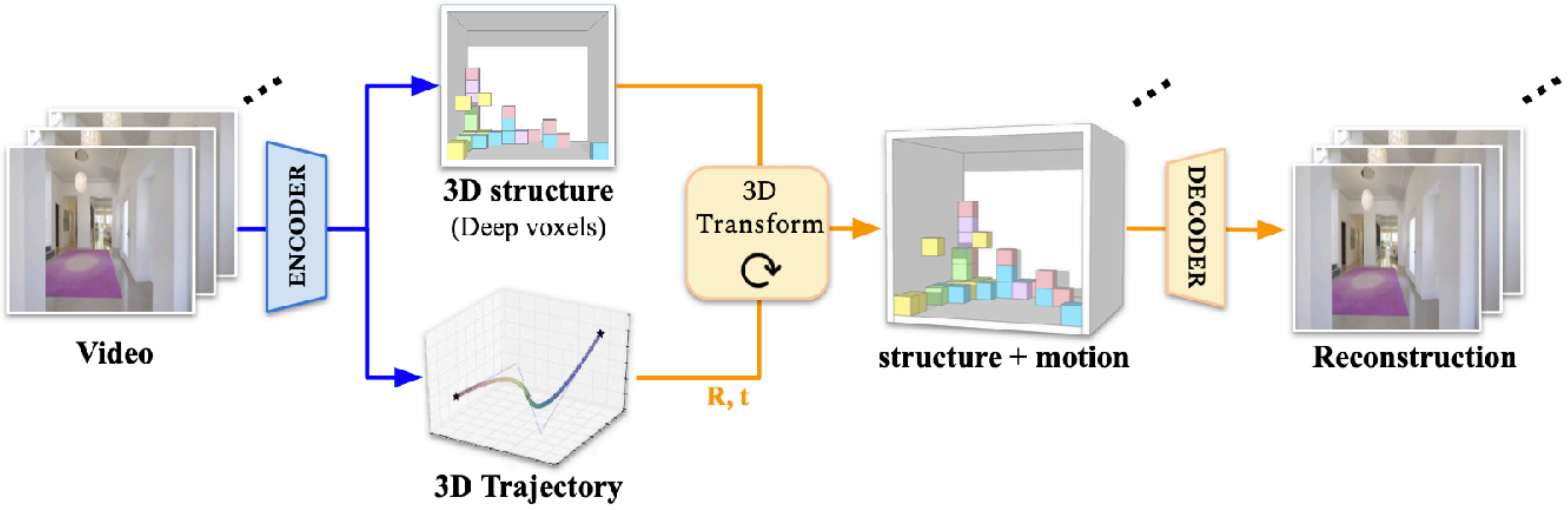
Video Autoencoder: self-supervised disentanglement of static 3D structure and motion



Zihang Lai
CMU

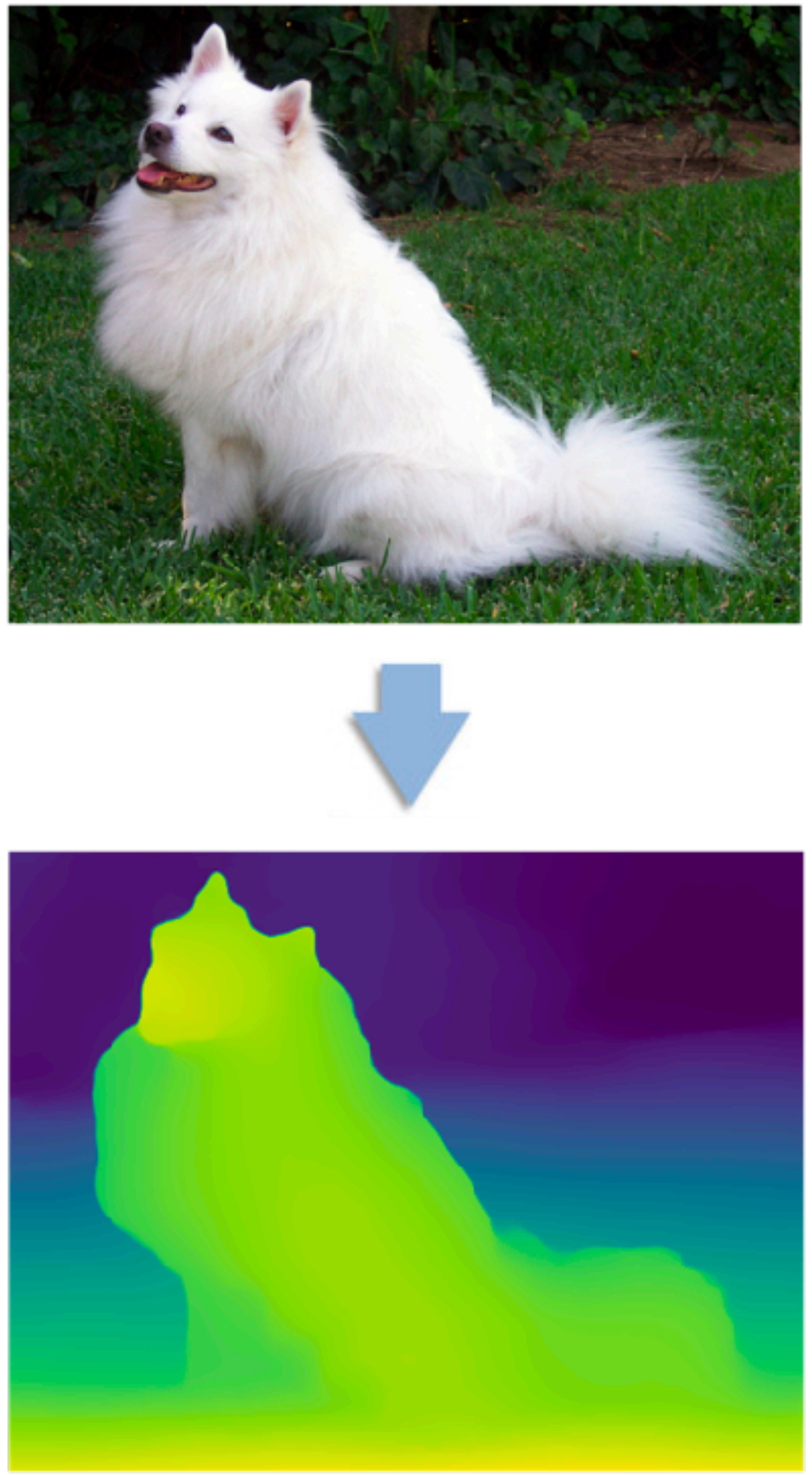
Sifei Liu
NVIDIA

Alexei A. Efros
UC Berkeley

Xiaolong Wang
UC San Diego

Disentangle the visual world

Movement

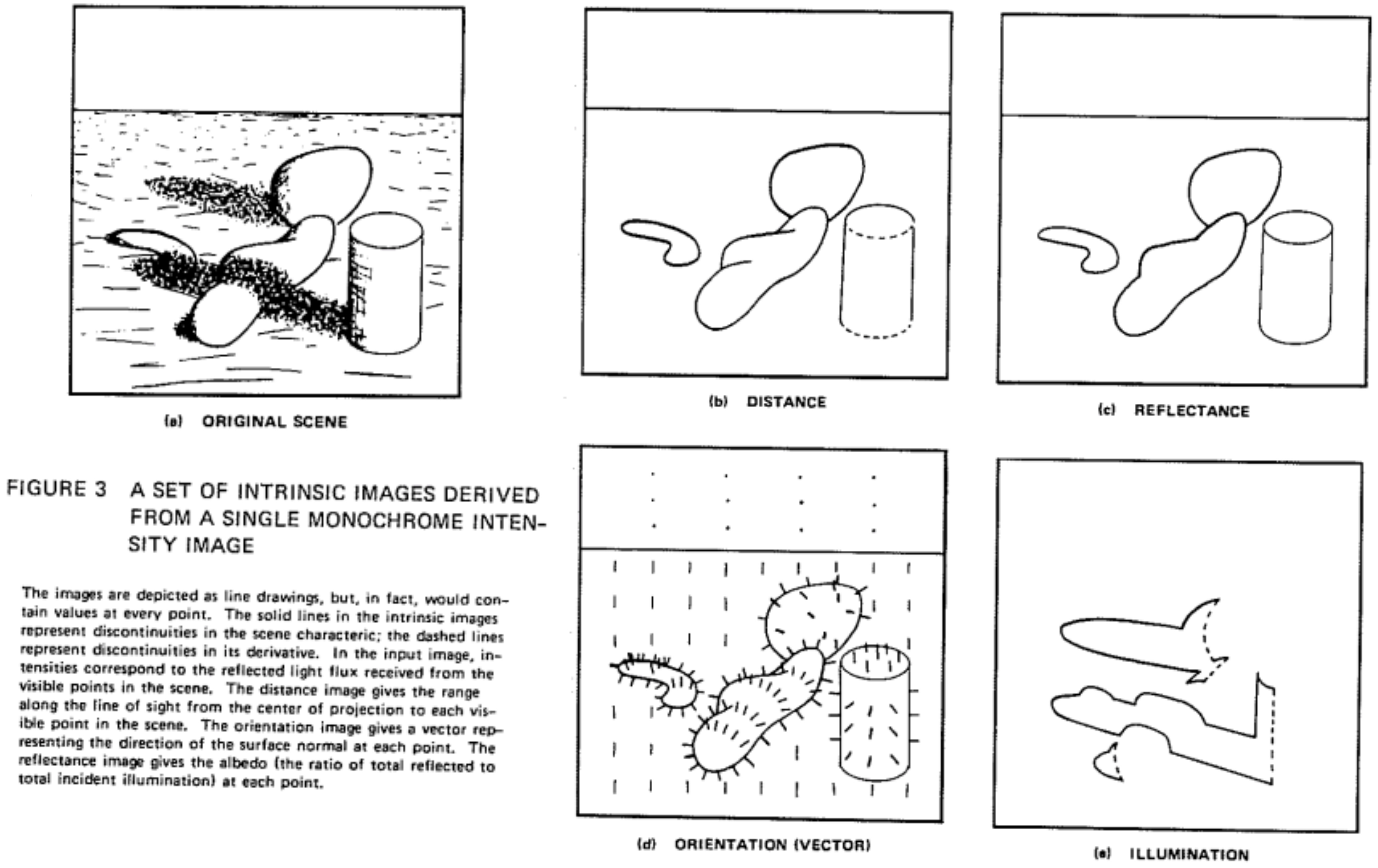


Depth

Structure and viewpoint

Prior work

From the very beginnings of computer vision, …



Barrow and Tenenbaum, Comput. Vis. Syst., 1978

$$
\begin{array}{r l}
 & \begin{array}{cccc} \text{Point 1} & \text{Point 2} & & \text{Point N} \end{array} \\
\begin{array}{r} \text{Image 1} \\ \text{Image 2} \\ \\ \text{Image F} \\ \text{Image 1} \\ \text{Image 2} \\ \\ \text{Image F} \end{array} &
\begin{bmatrix}
\tilde{u}_{1,1} & \tilde{u}_{1,2} & \cdots & \tilde{u}_{1,N} \\
\tilde{u}_{2,1} & \tilde{u}_{2,2} & \cdots & \tilde{u}_{2,N} \\
\vdots & \vdots & \vdots & \vdots \\
\tilde{u}_{F,1} & \tilde{u}_{F,2} & \cdots & \tilde{u}_{F,N} \\
\tilde{v}_{1,1} & \tilde{v}_{1,2} & \cdots & \tilde{v}_{1,N} \\
\tilde{u}_{2,1} & \tilde{u}_{2,2} & \cdots & \tilde{v}_{2,N} \\
\vdots & \vdots & \vdots & \vdots \\
\tilde{v}_{F,1} & \tilde{v}_{F,2} & \cdots & \tilde{v}_{F,N}
\end{bmatrix}
\end{array}
=
\begin{bmatrix}
\mathbf{i}_1^T \\ \mathbf{i}_2^T \\ \vdots \\ \mathbf{i}_F^T \\ \mathbf{j}_1^T \\ \mathbf{j}_2^T \\ \vdots \\ \mathbf{j}_F^T
\end{bmatrix}
\begin{array}{c}
\begin{array}{cccc} \text{Point 1} & \text{Point 2} & & \text{Point N} \end{array} \\
[P_1 \quad P_2 \quad \cdots \quad P_N]
\end{array}
$$

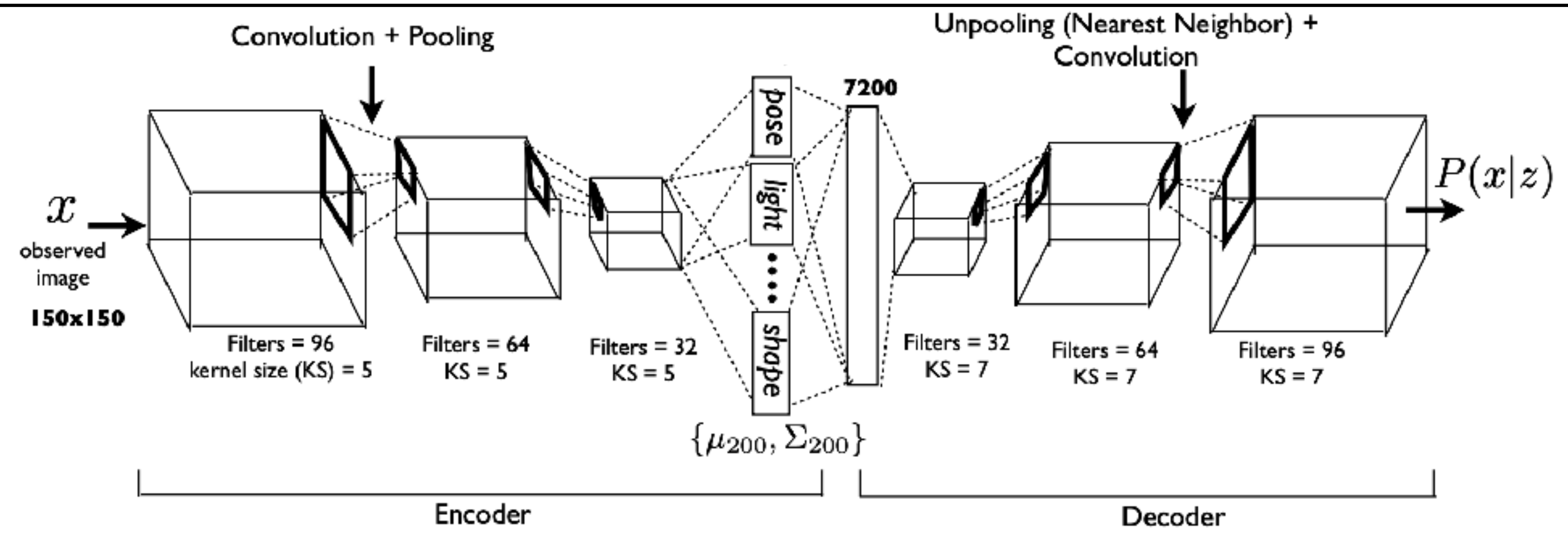
Scene structure

Measurement matrix Camera motion

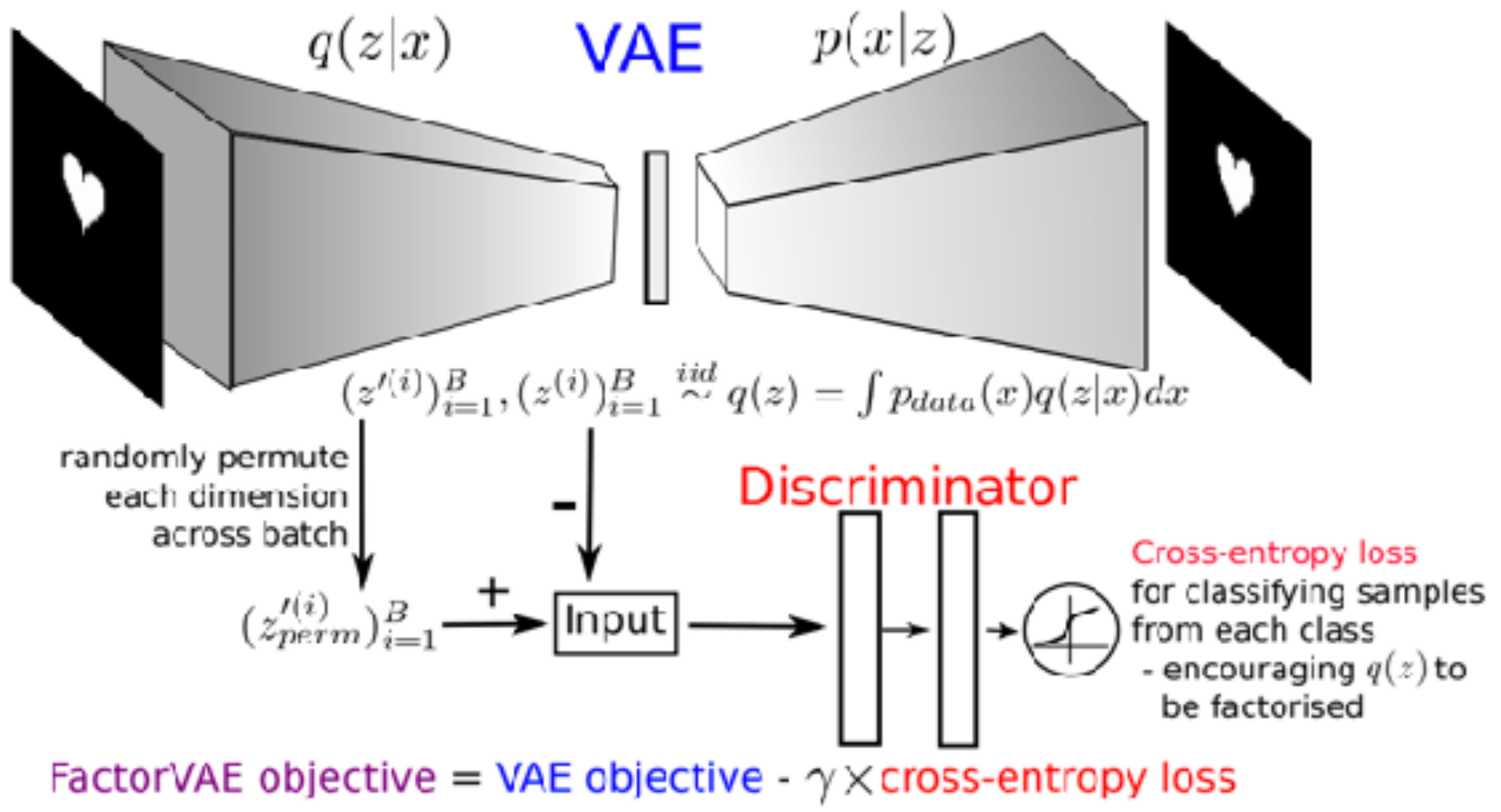
Tomasi and Kanade, *IJCV*, 1992

Prior work

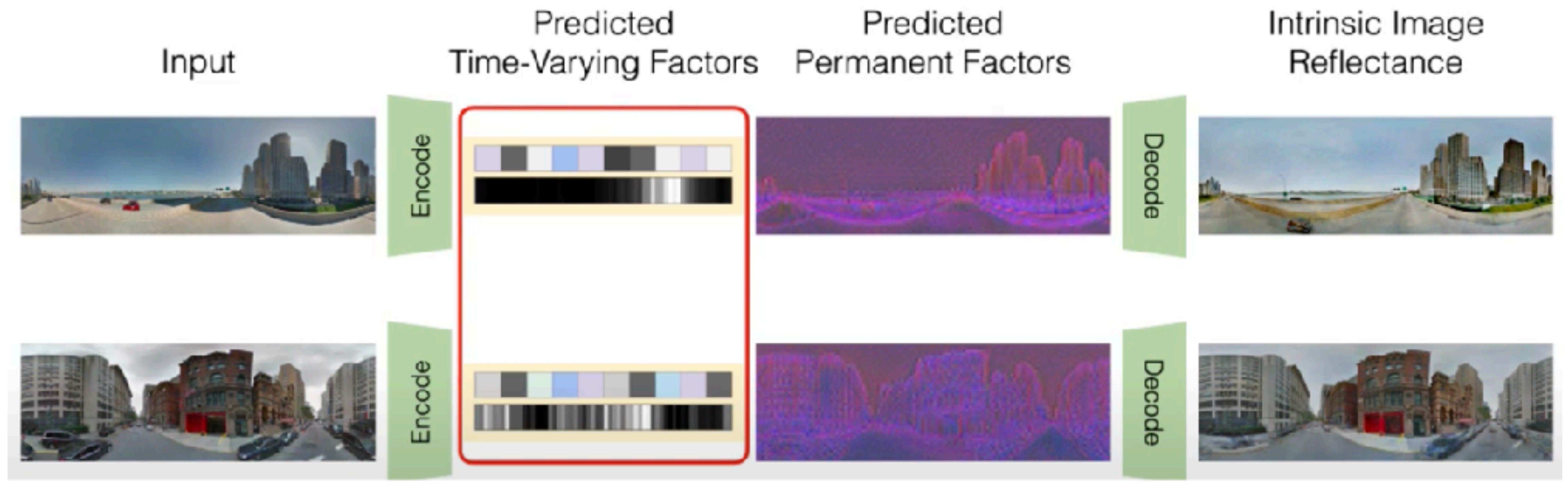
Learning disentangled visual representation from auto-encoders



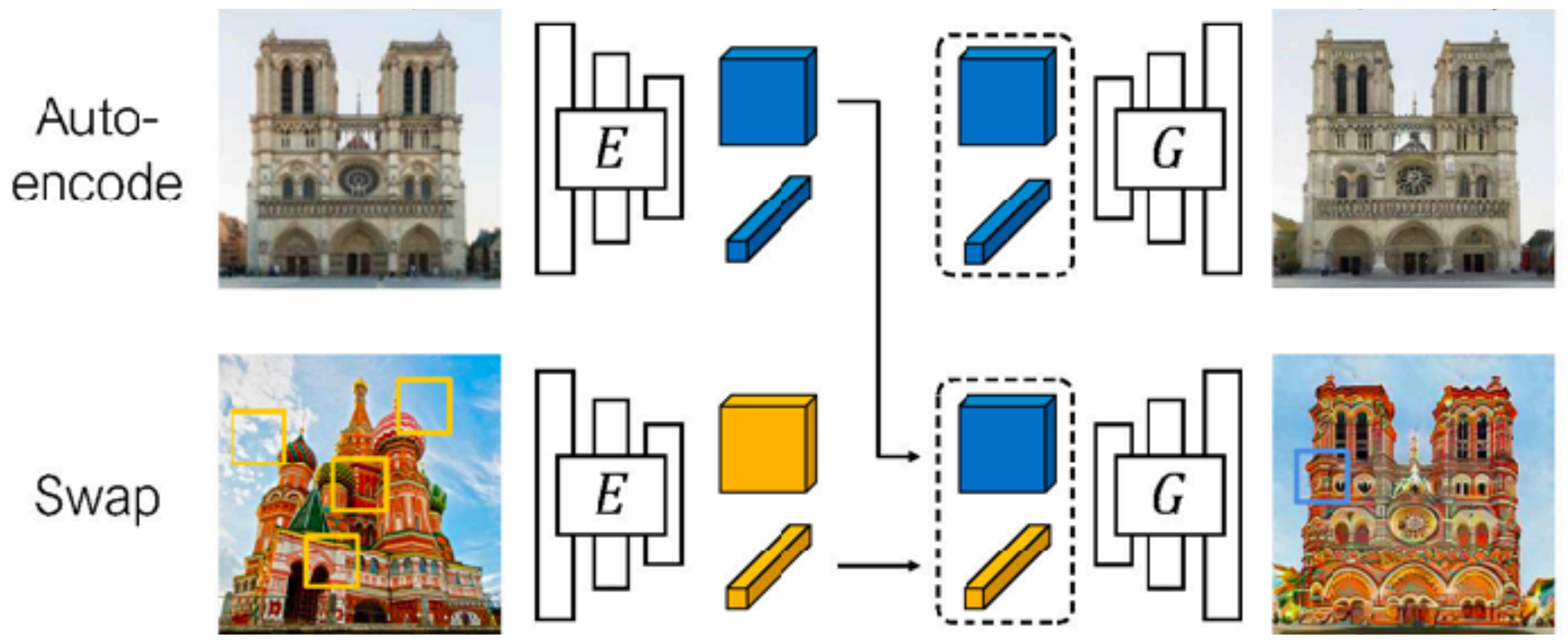
Kulkarni et al., NIPS, 2015



Kim et al., ICML, 2018



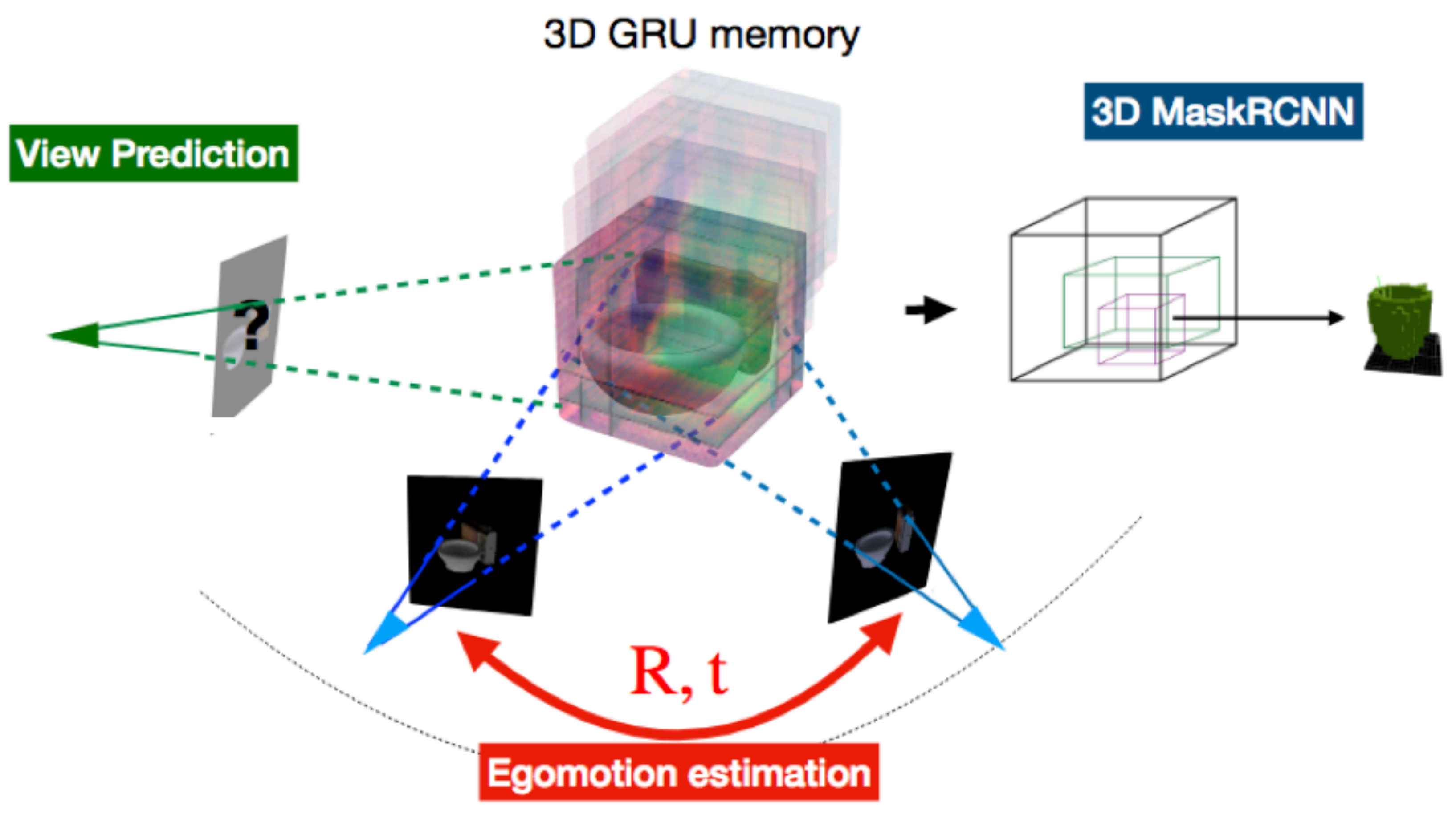
Liu et al., ECCV, 2020



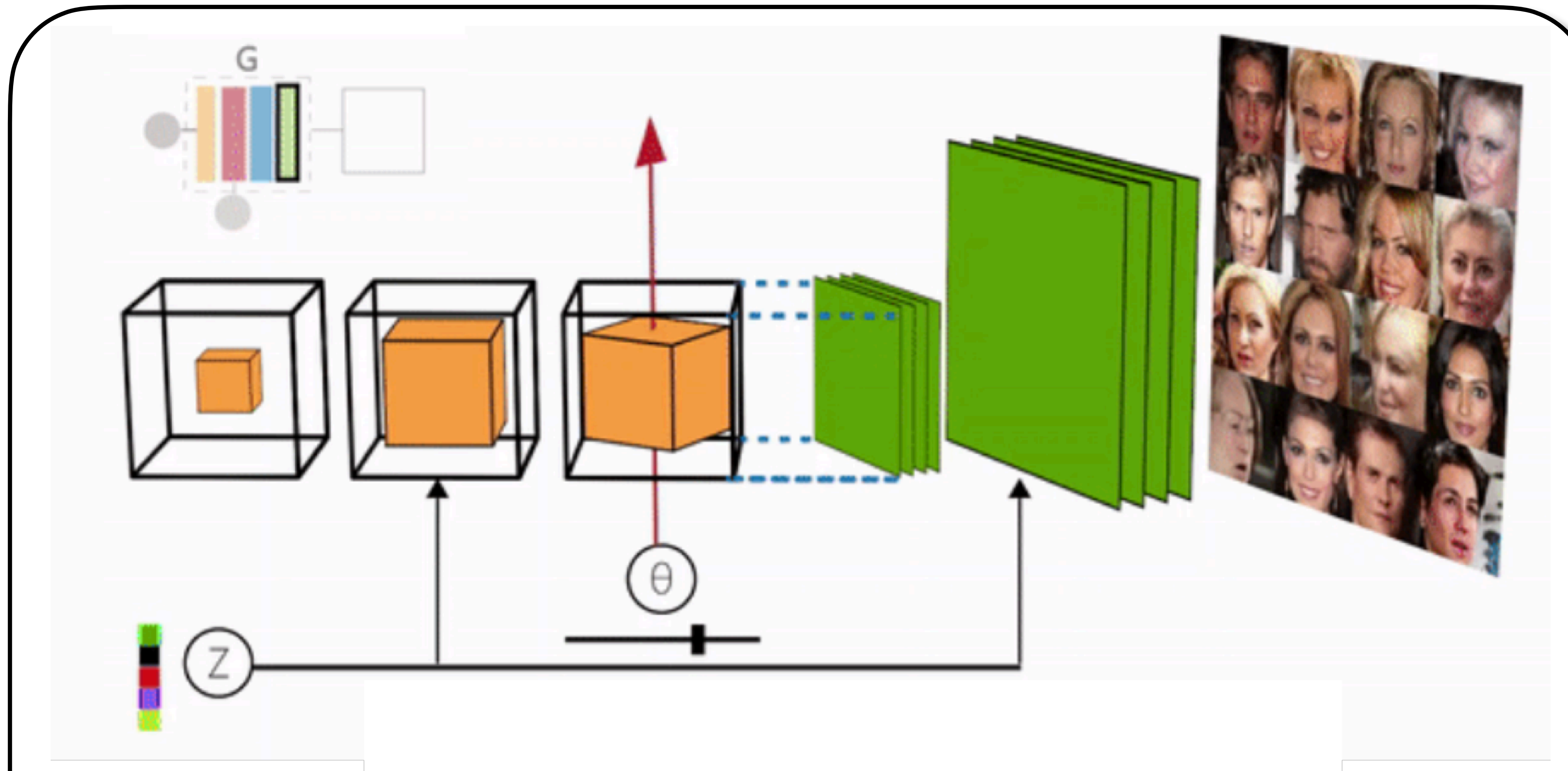
Park et al., NeurIPS, 2020

Prior work

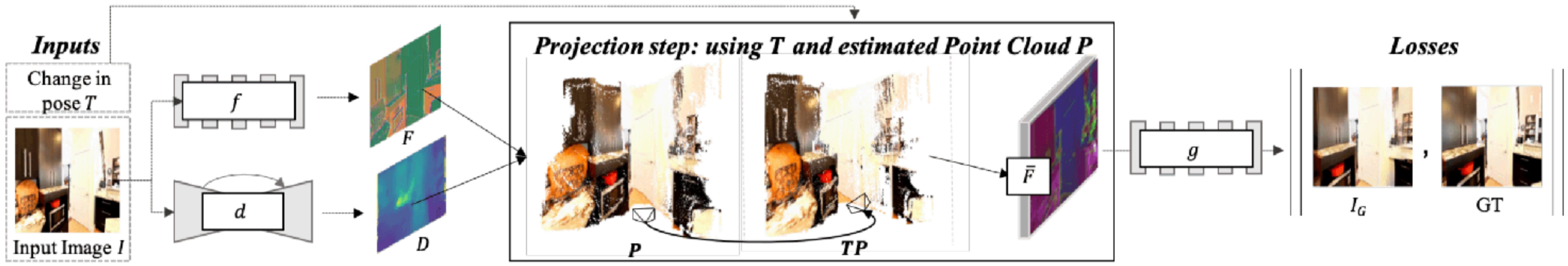
Learning 3D from 2D supervisions



Tung et al., *CVPR,* 2019



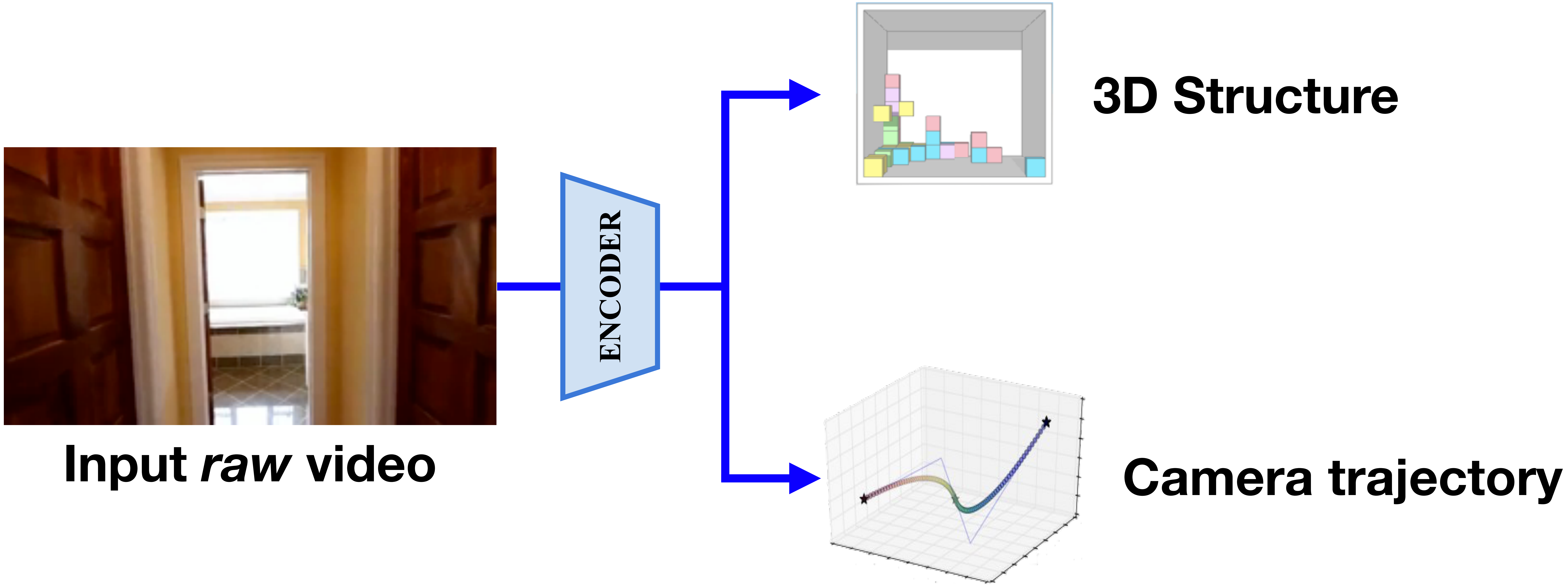
Nguyen-Phuoc et al., *ICCV,* 2019



Wiles et al., *CVPR,* 2020

Objective

In this work, we learn to separate 3D structure from Camera Motion without any human annotations



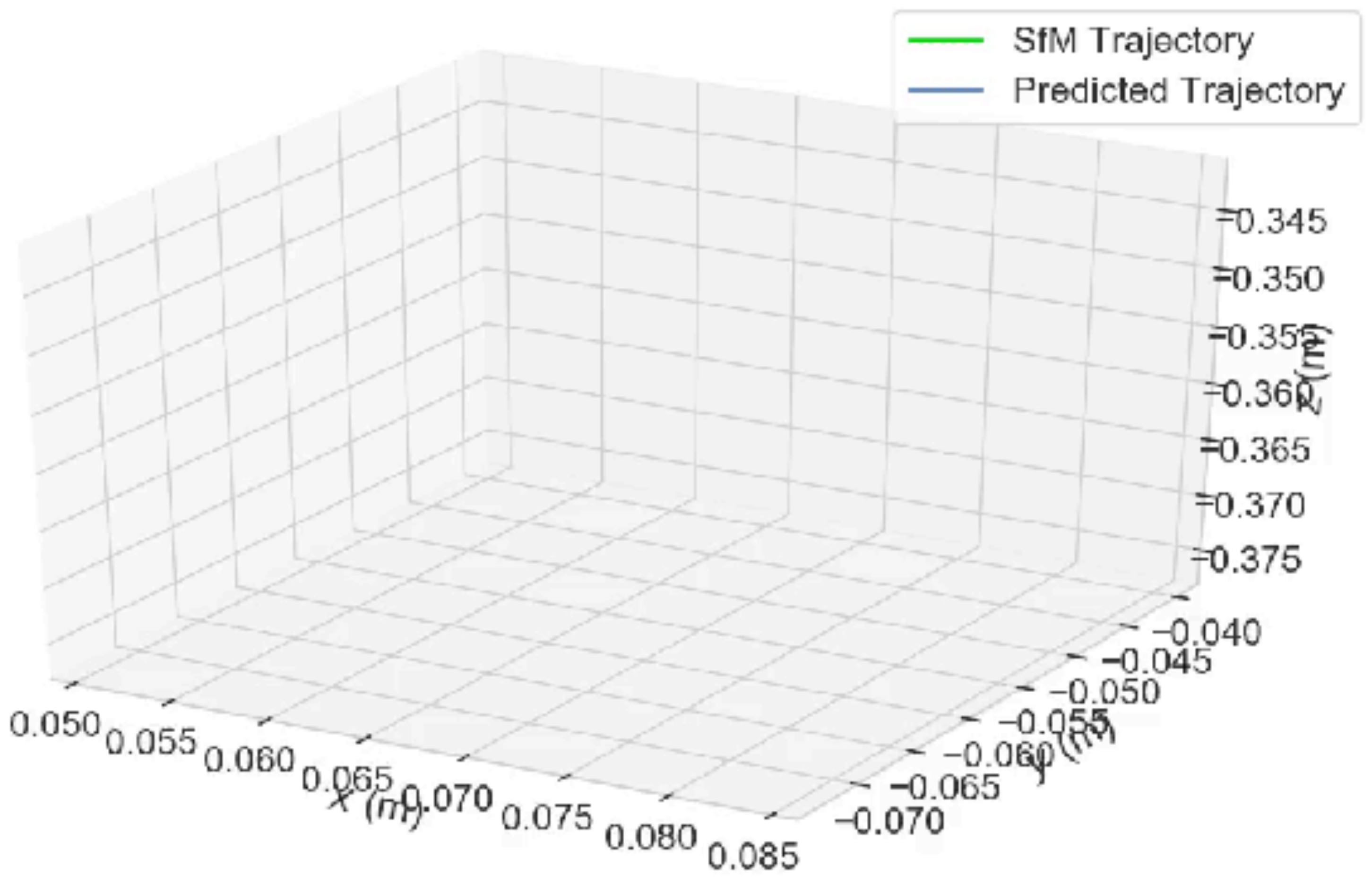
Test time

The features obtained (3D structure, Camera Motion) can be used for several downstream tasks:

Single image

Novel view synthesis

Raw video



Pose estimation in video

Single image + Raw video

Video following

*Actual results

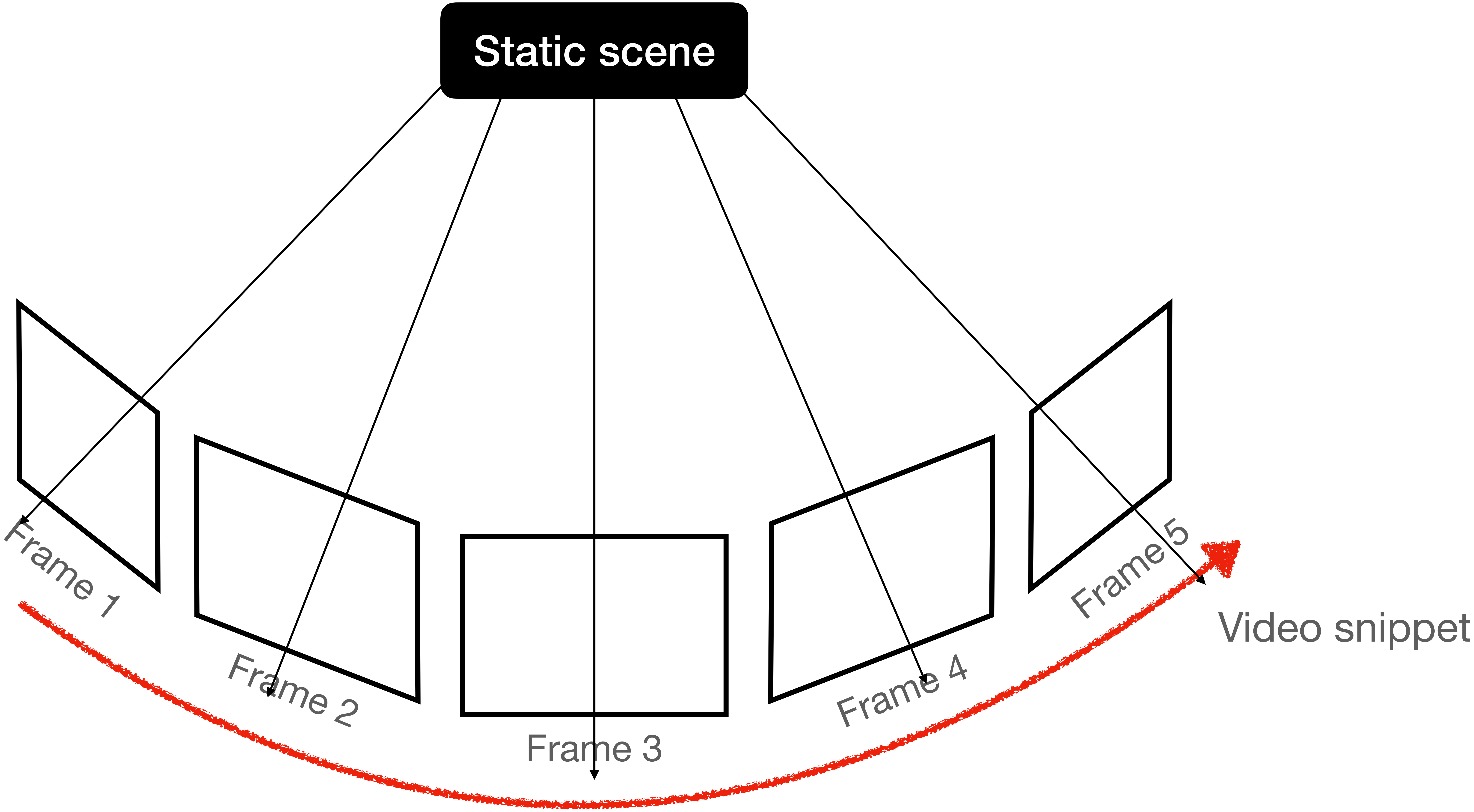
Learning from temporal continuity of videos

Spatio-temporal continuity

No spatio-temporal continuity

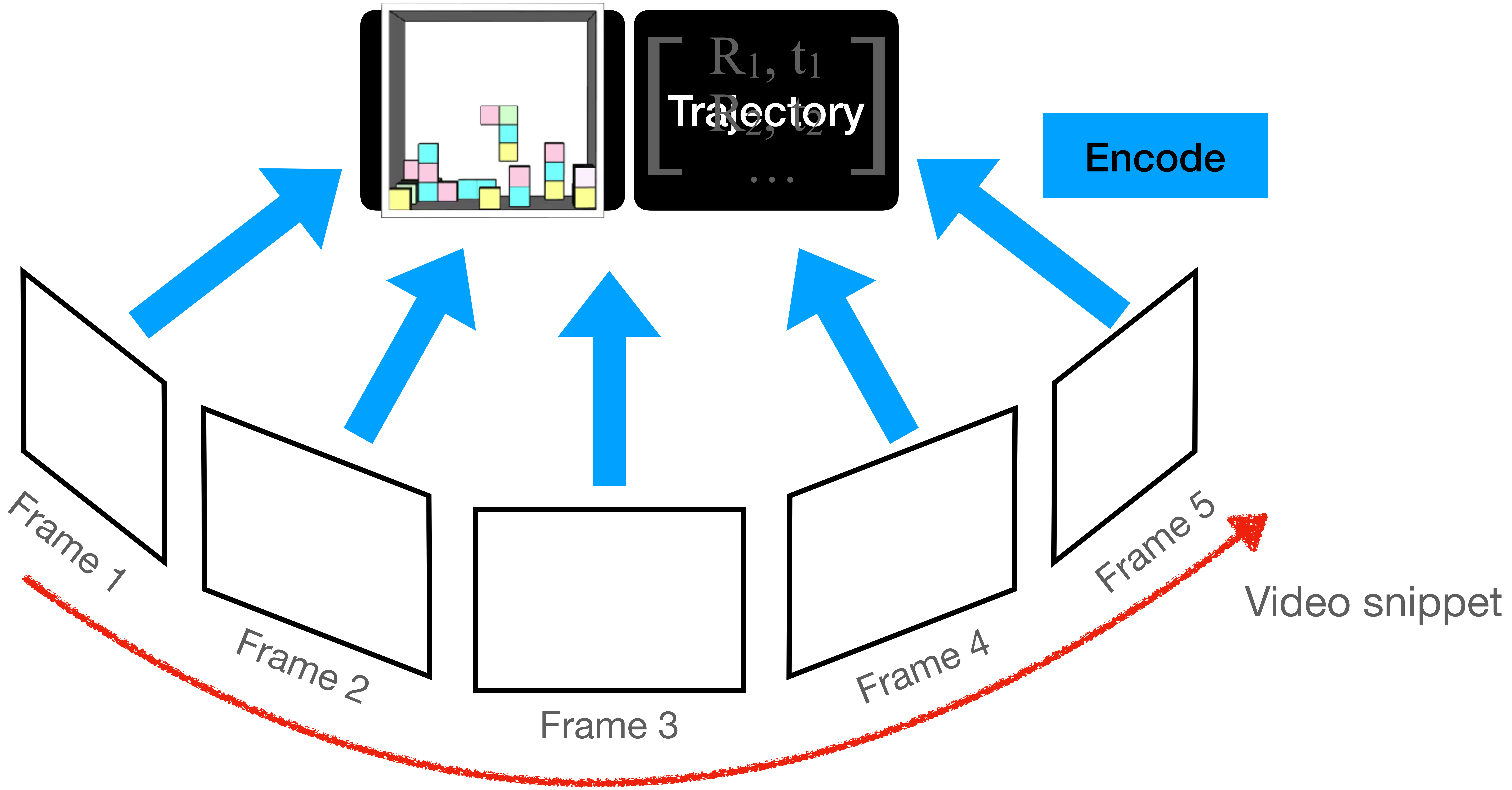
Learning from temporal continuity of videos

Assume that a local snippet of video is capturing a static scene



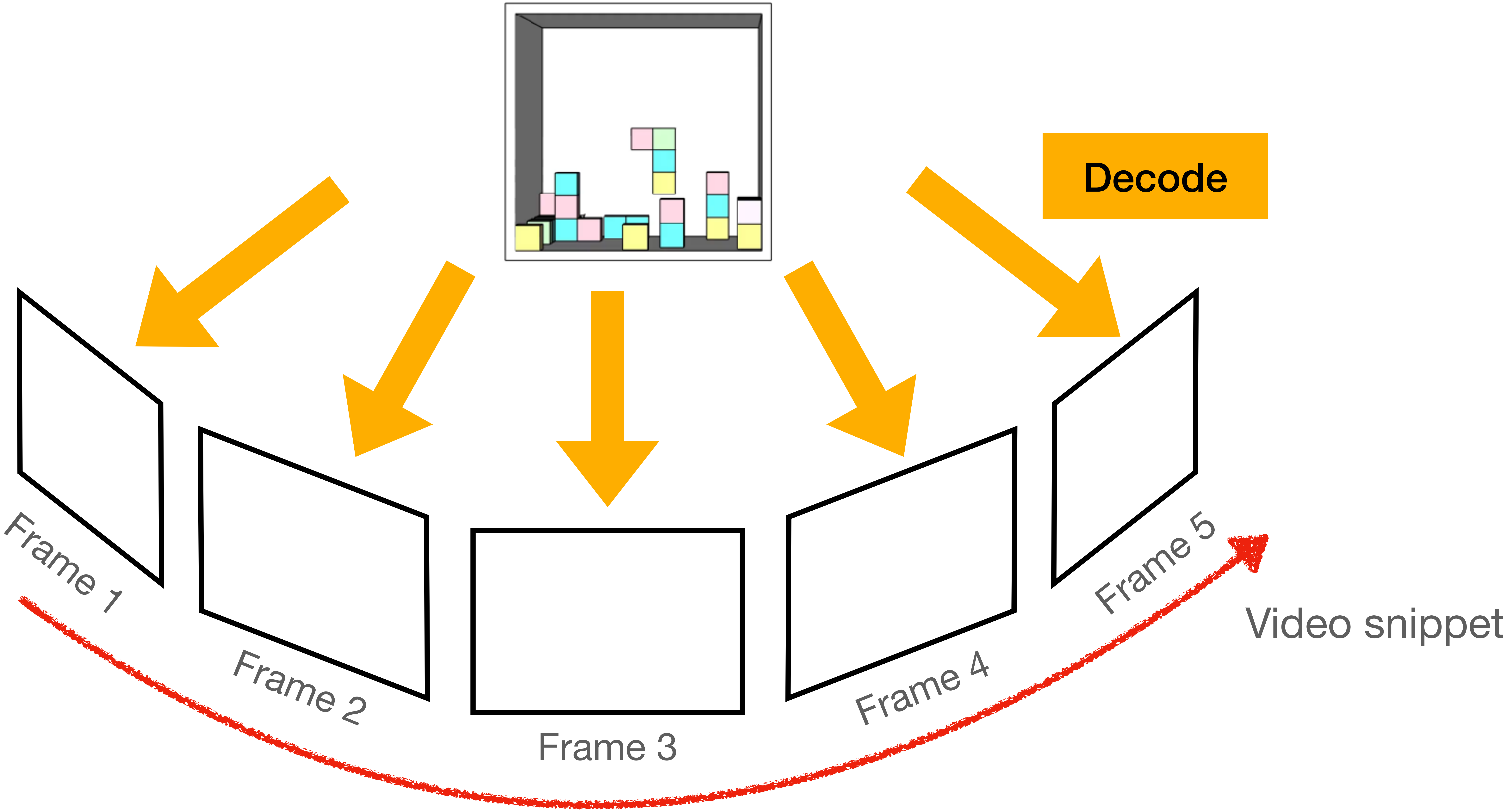
Learning from temporal continuity of videos

Assume that a local snippet of video is capturing a static scene

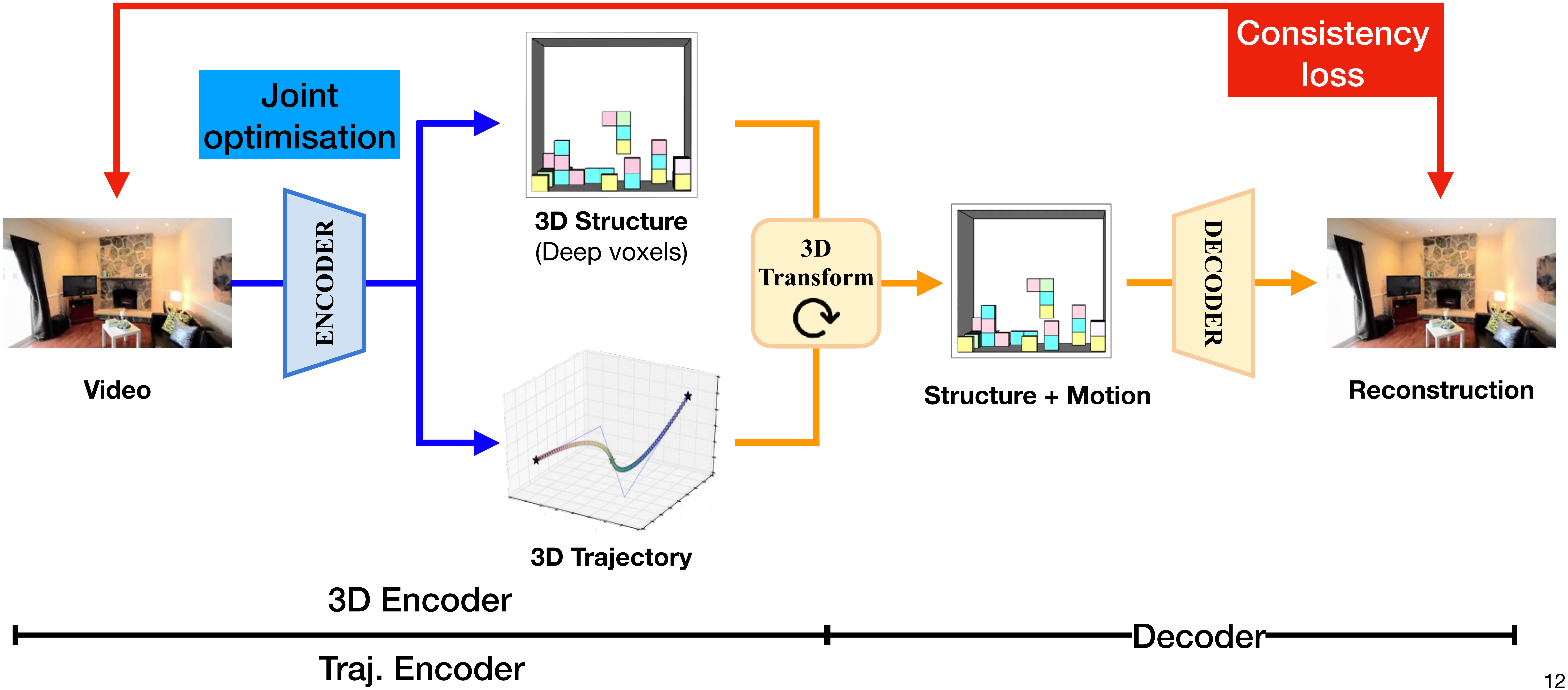


Learning from temporal continuity of videos

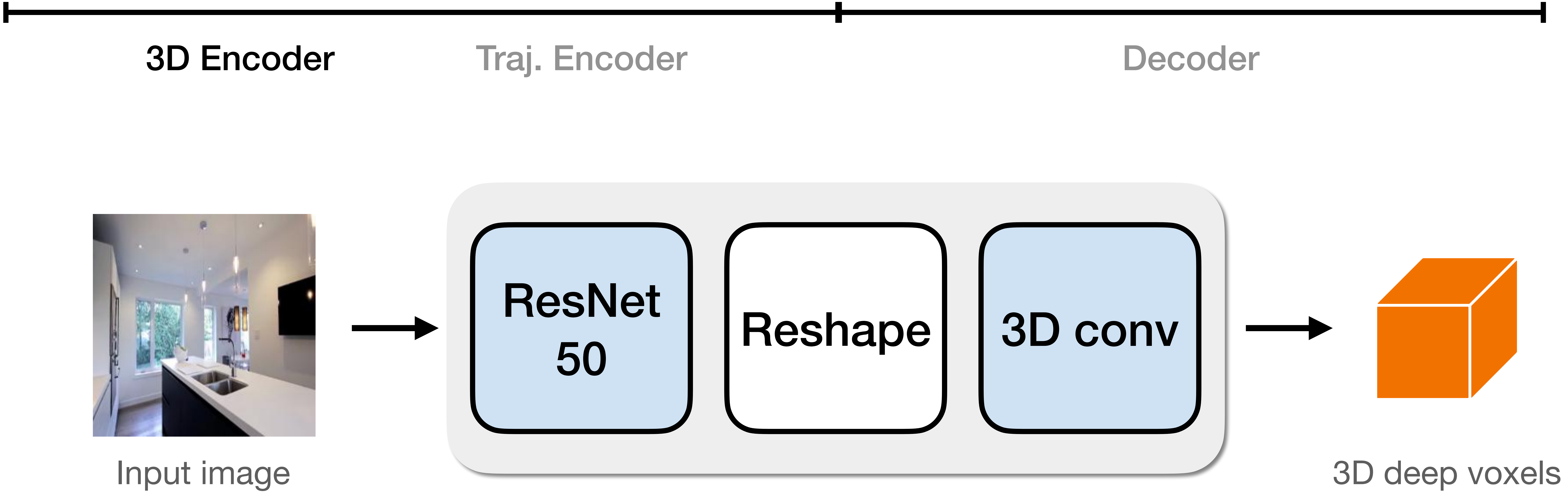
Assume that a local snippet of video is capturing a static scene



Model architecture



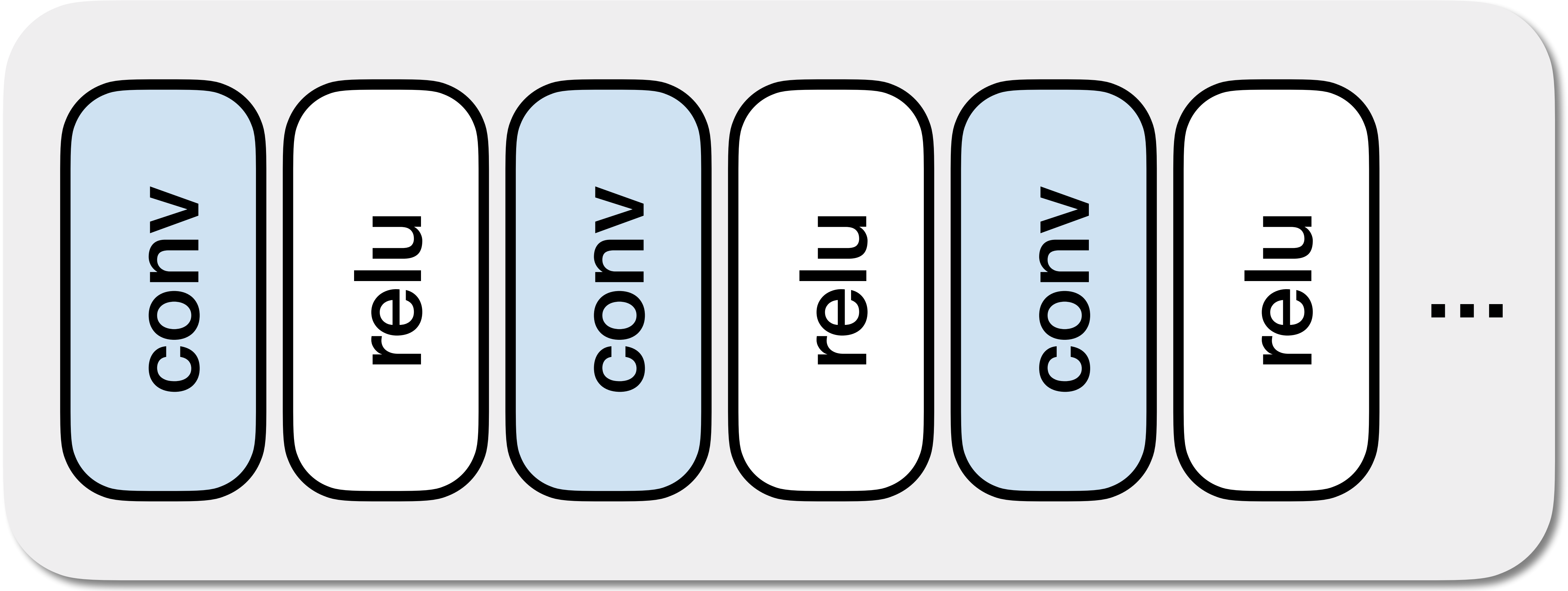
Model architecture



Model architecture

3D Encoder Traj. Encoder Decoder

Stacked image pair

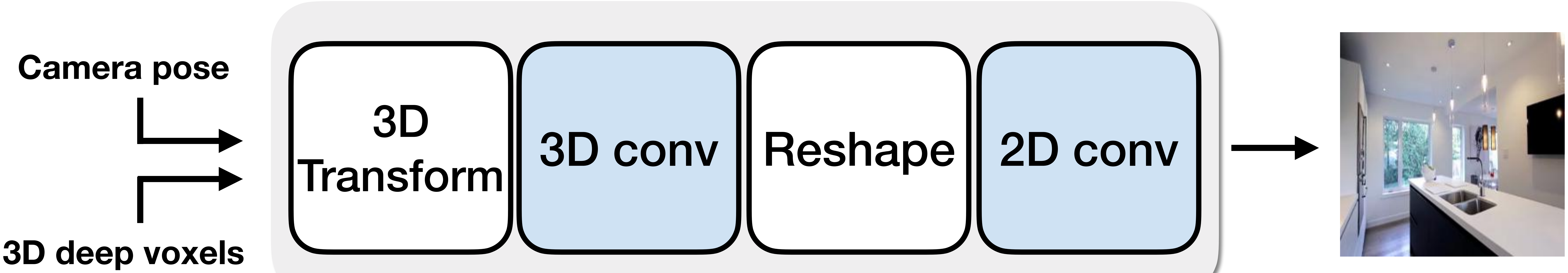


$[R, t]$

6D Pose

Model architecture

3D Encoder Traj. Encoder **Decoder**



Output image

Training loss

- Reconstruction loss:
  - $L_r(I_t, \hat{I}_t) = \|I_t - \hat{I}_t\|_1 + L_p(I_t, \hat{I}_t)$
- GAN loss:
  - $L_g(\hat{I}_t) = -F_D(\hat{I})$
- Consistency loss between deep voxels:
  - $L_c(V_{t1}, V_{t2}) = \|V_{t1} - R(V_{t2}, P_{t2 \to t1})\|_1$

Video Autoencoder

Results

Datasets

RealEstate10K

Matterport3D

Replica

Results

Novel view synthesis

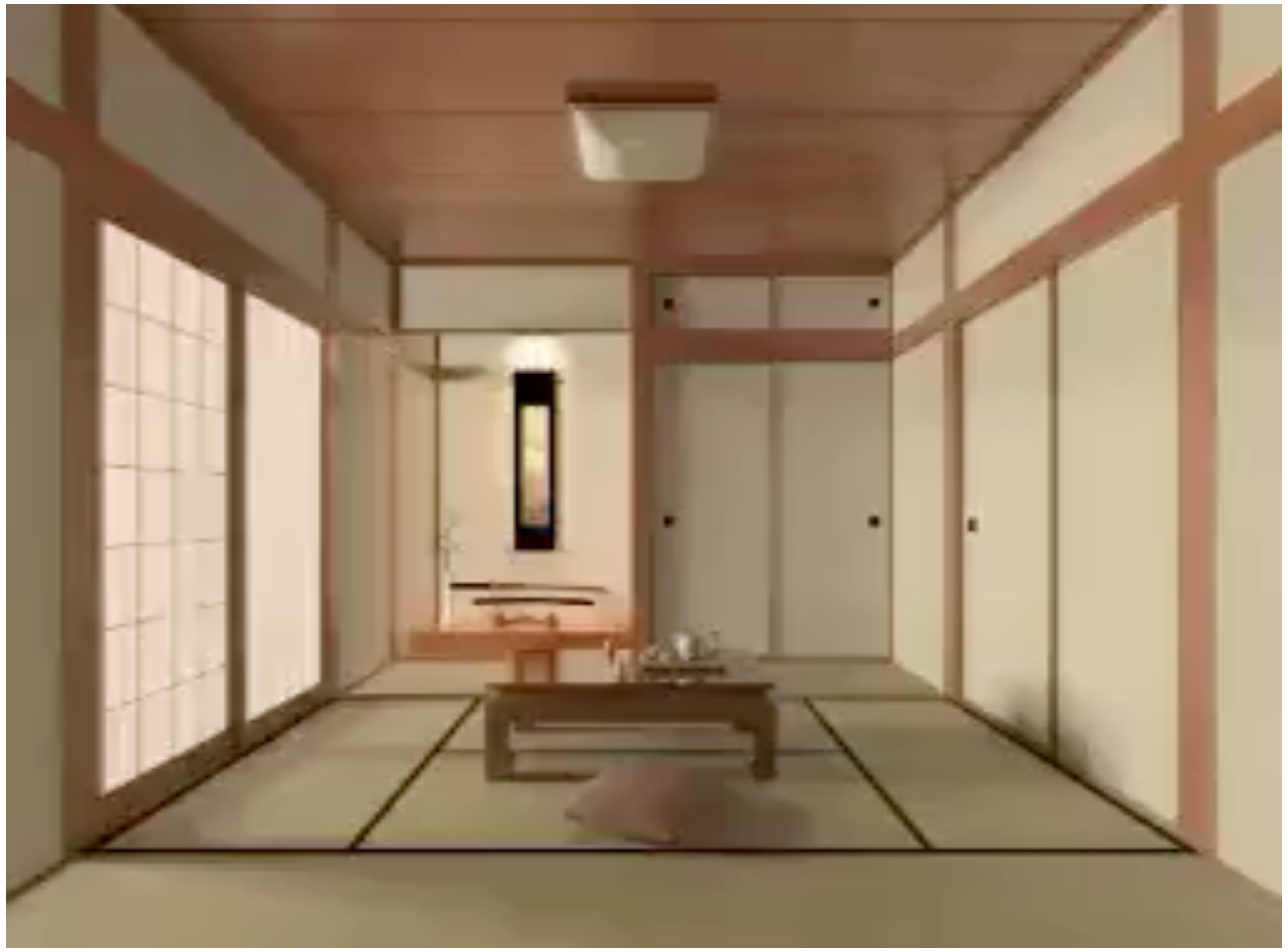
Single input image | **Output video**

RealEstate10K dataset

Results

Novel view synthesis (Out-of-domain results)

Single input image **Output video**



A Japanese living room (out-of-distribution)

Results

Novel view synthesis (Out-of-domain results)

Single input image

Output video

“Spirited Away”

Results

Novel view synthesis (Out-of-domain results)

Single input image

Output video

Bedroom in Arles, Vincent van Gogh

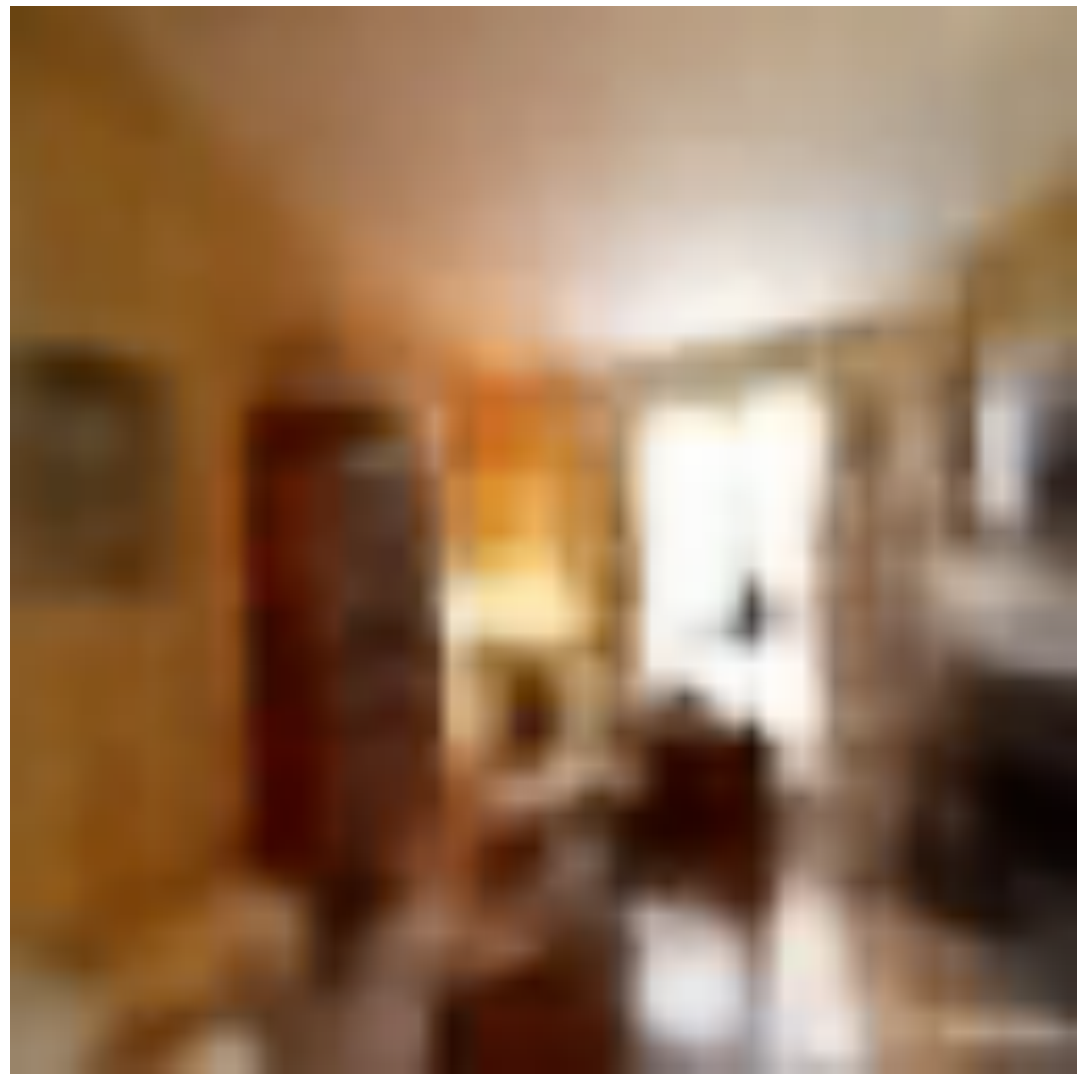
Results

Comparison with previous methods

Input image

RealEstate10K dataset

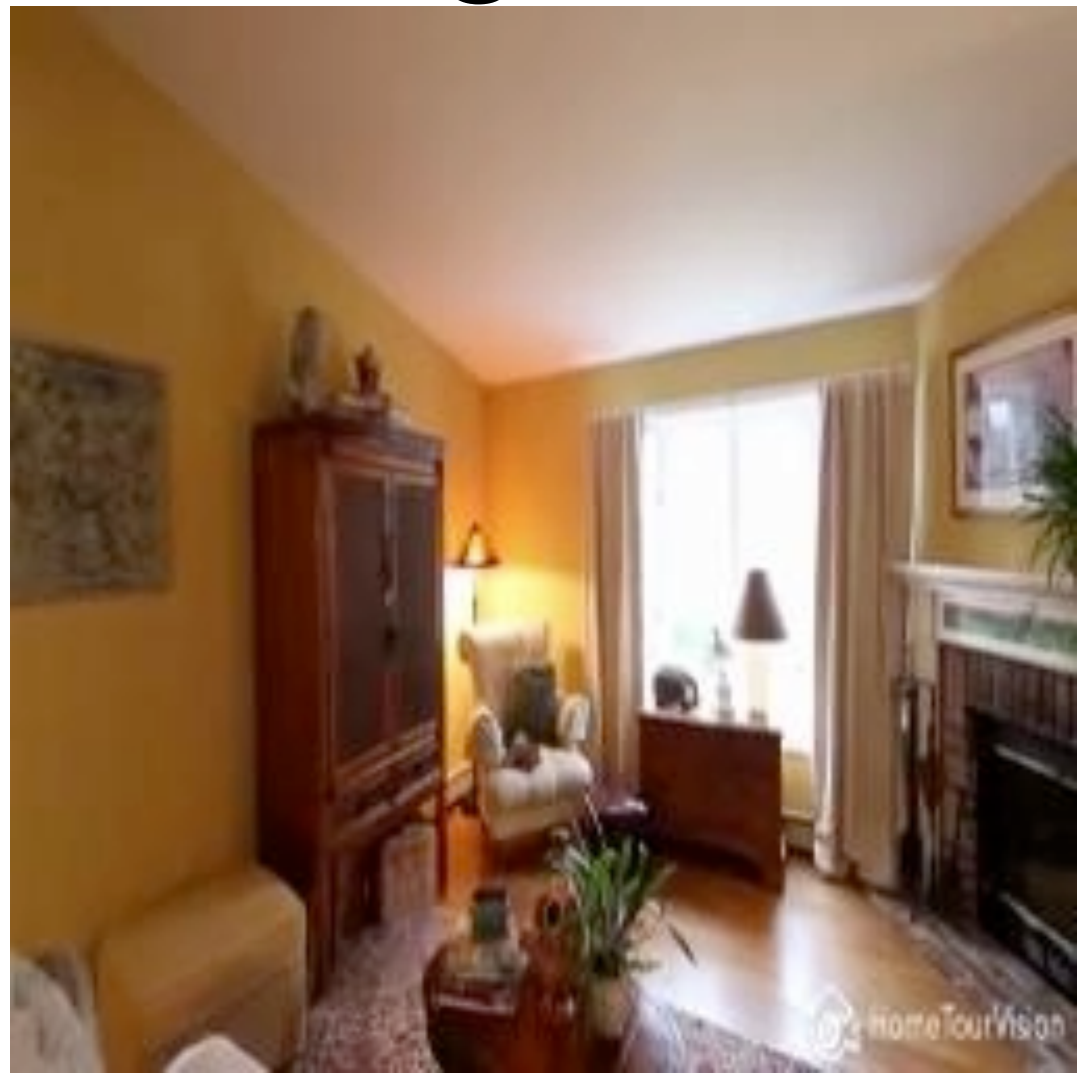
Mustikovela et al.



Tung et al.

Yu et al.

Wiles et al.



Ours

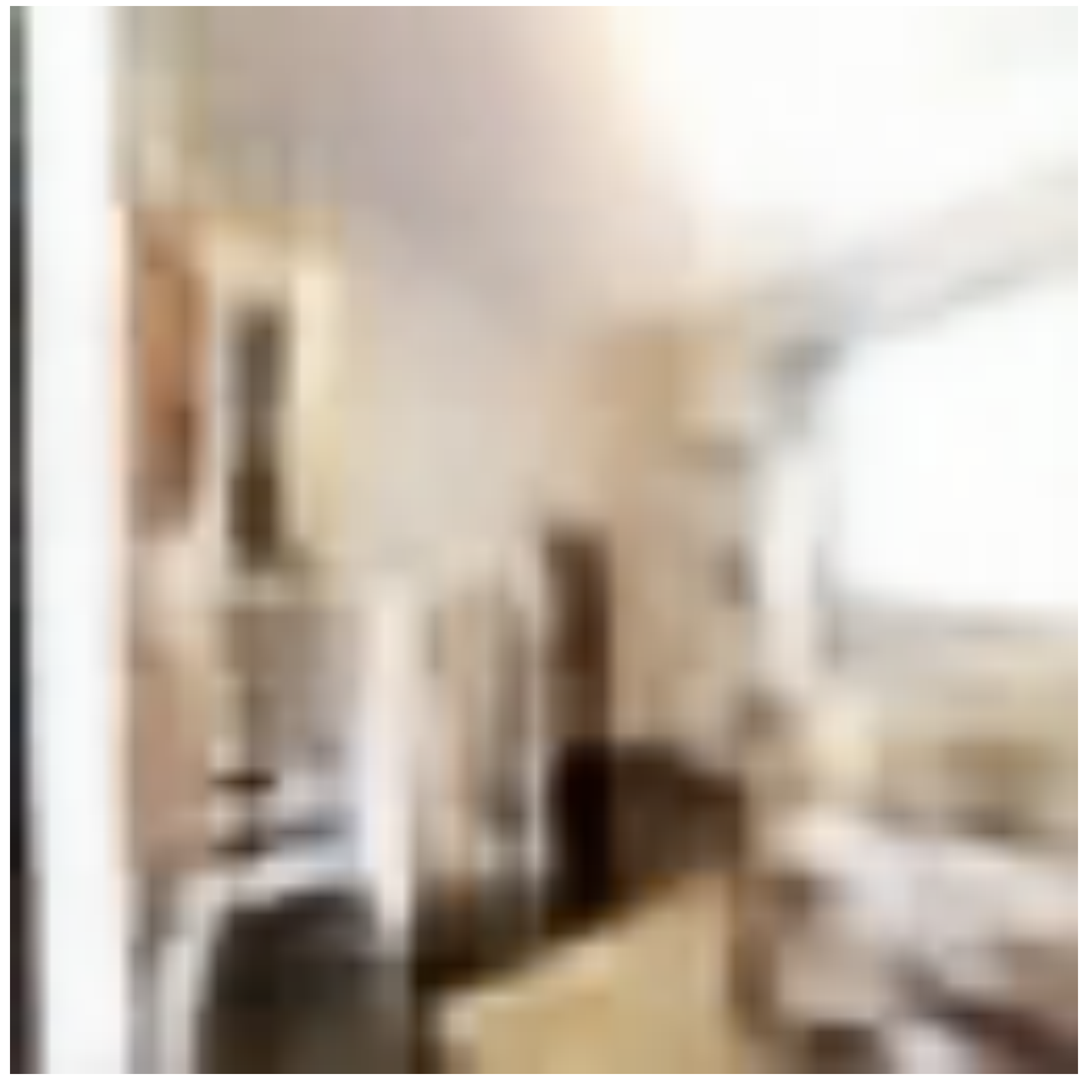
Ground truth

Results **Comparison with previous methods**

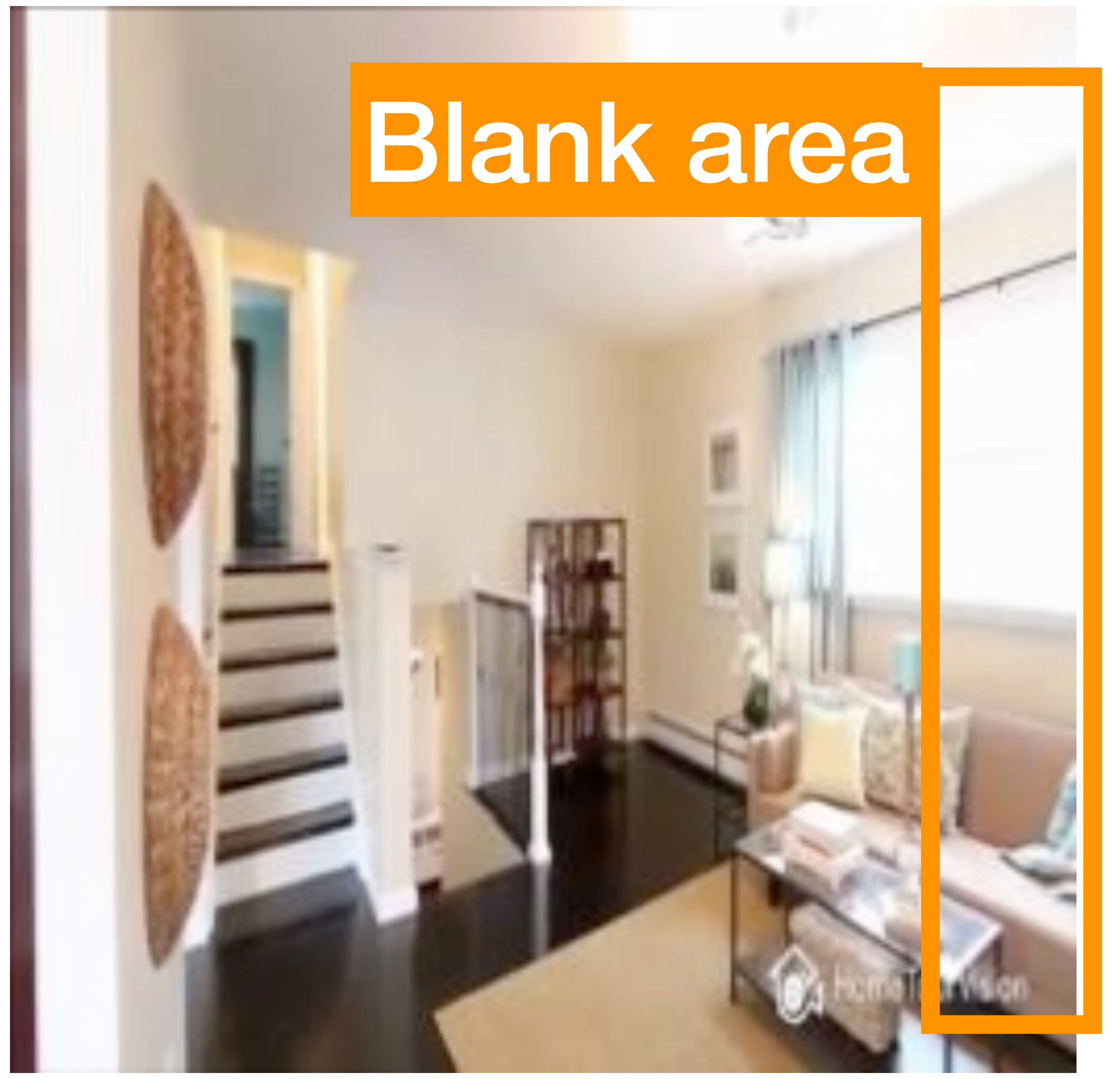
Input image

RealEstate10K dataset

Mustikovela et al.

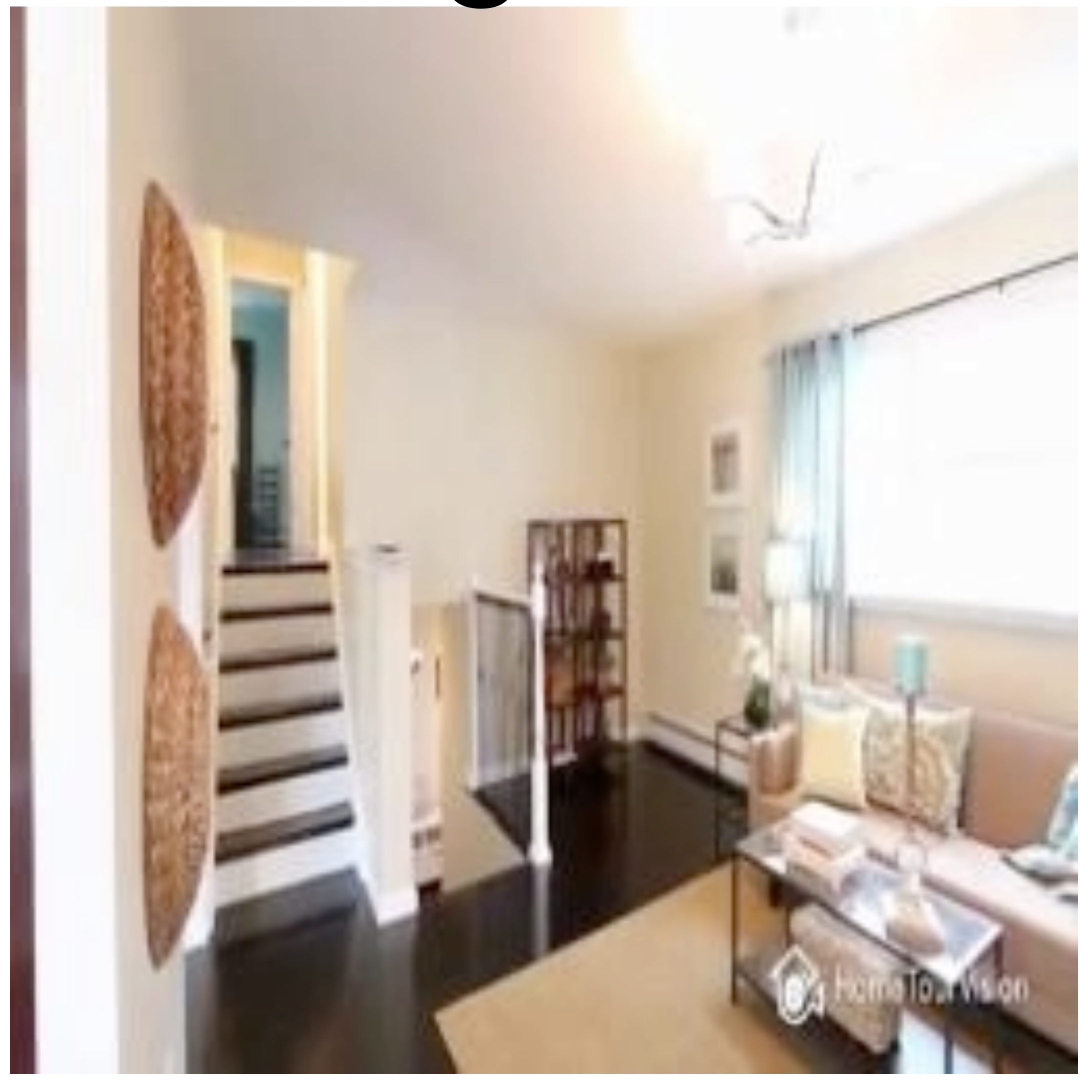


Tung et al.

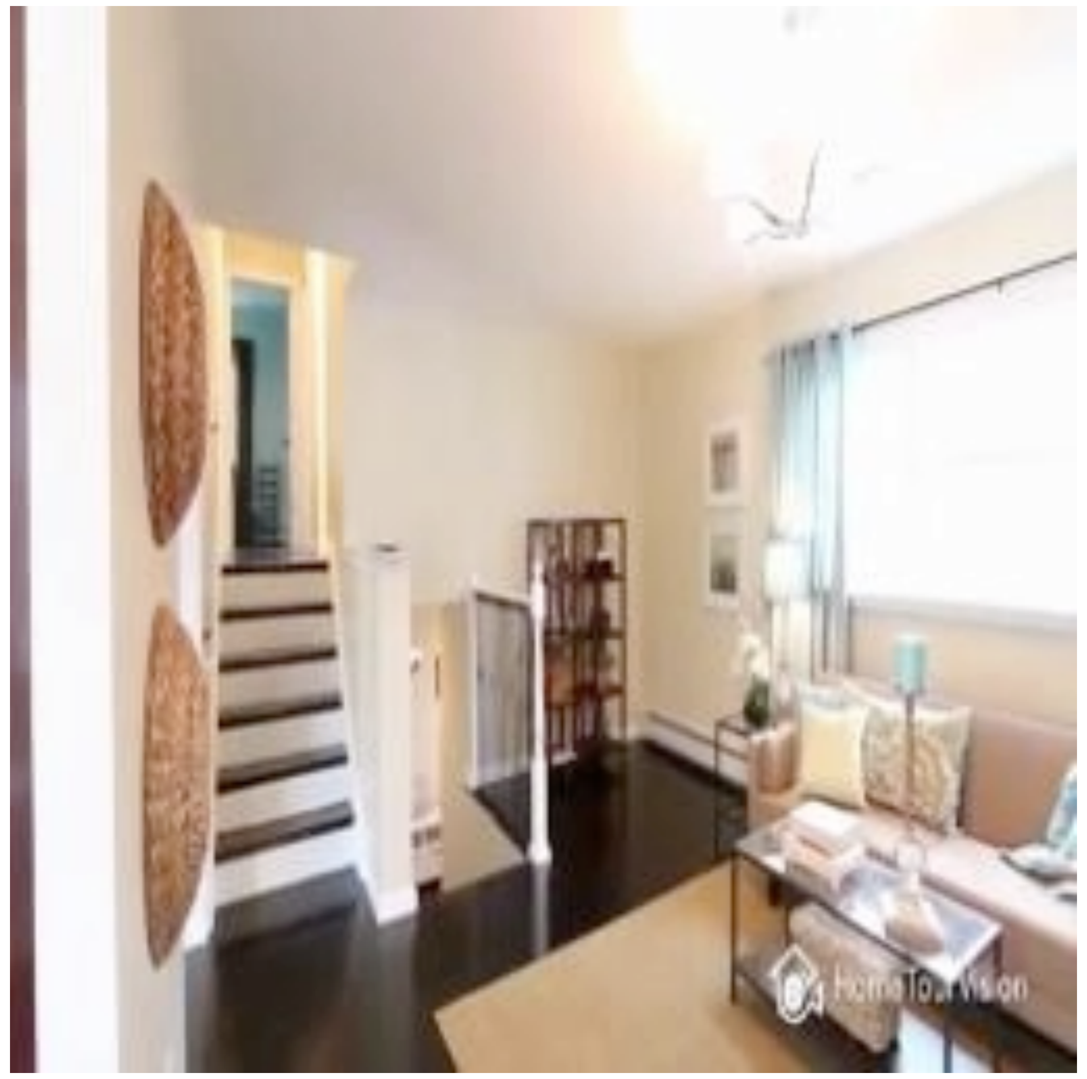


Yu et al.

Wiles et al.



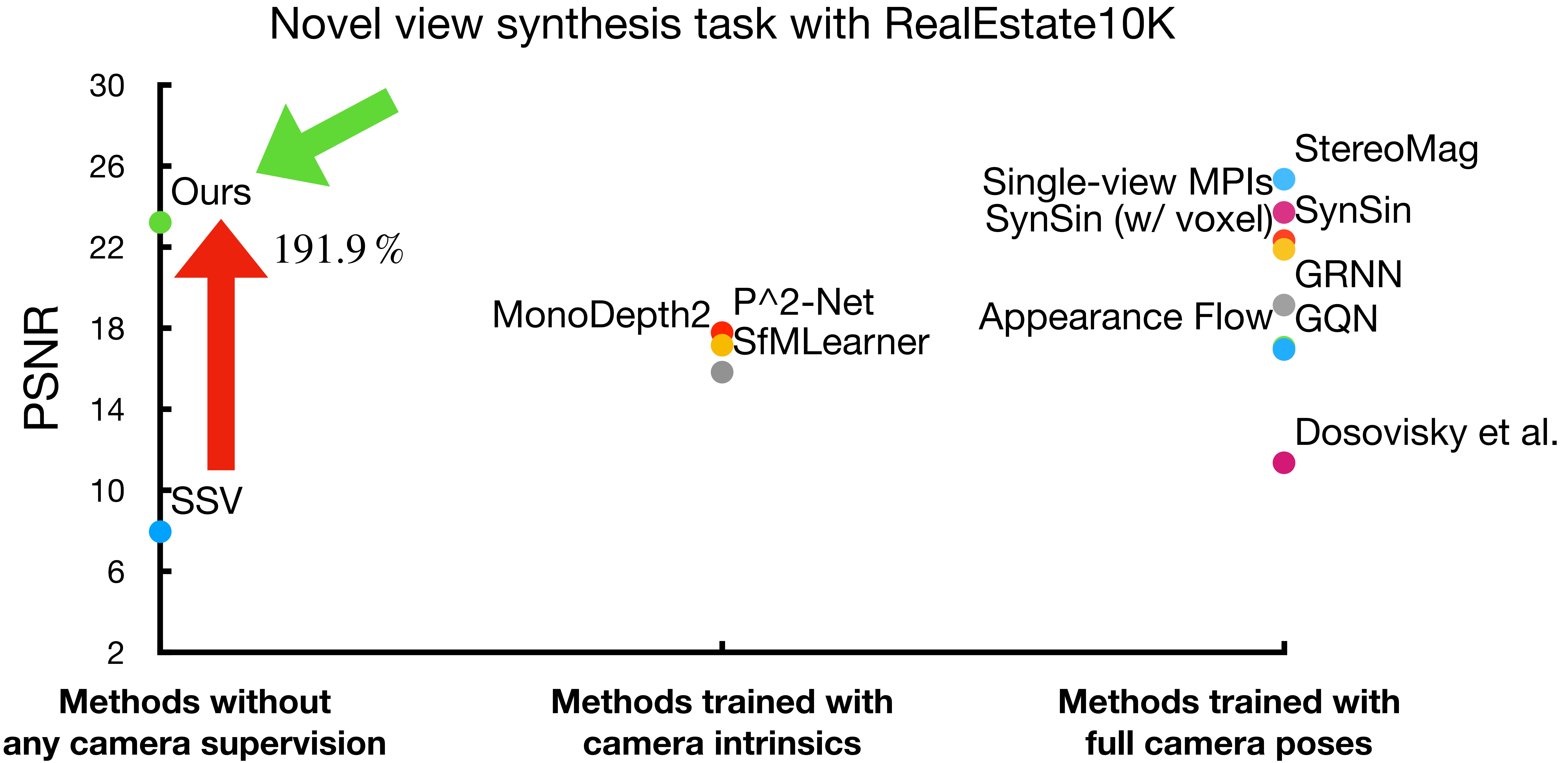
Ours



Ground truth

Results

Novel view synthesis (RealEstate10K)



Results

Novel view synthesis (Matterport3D & Replica)

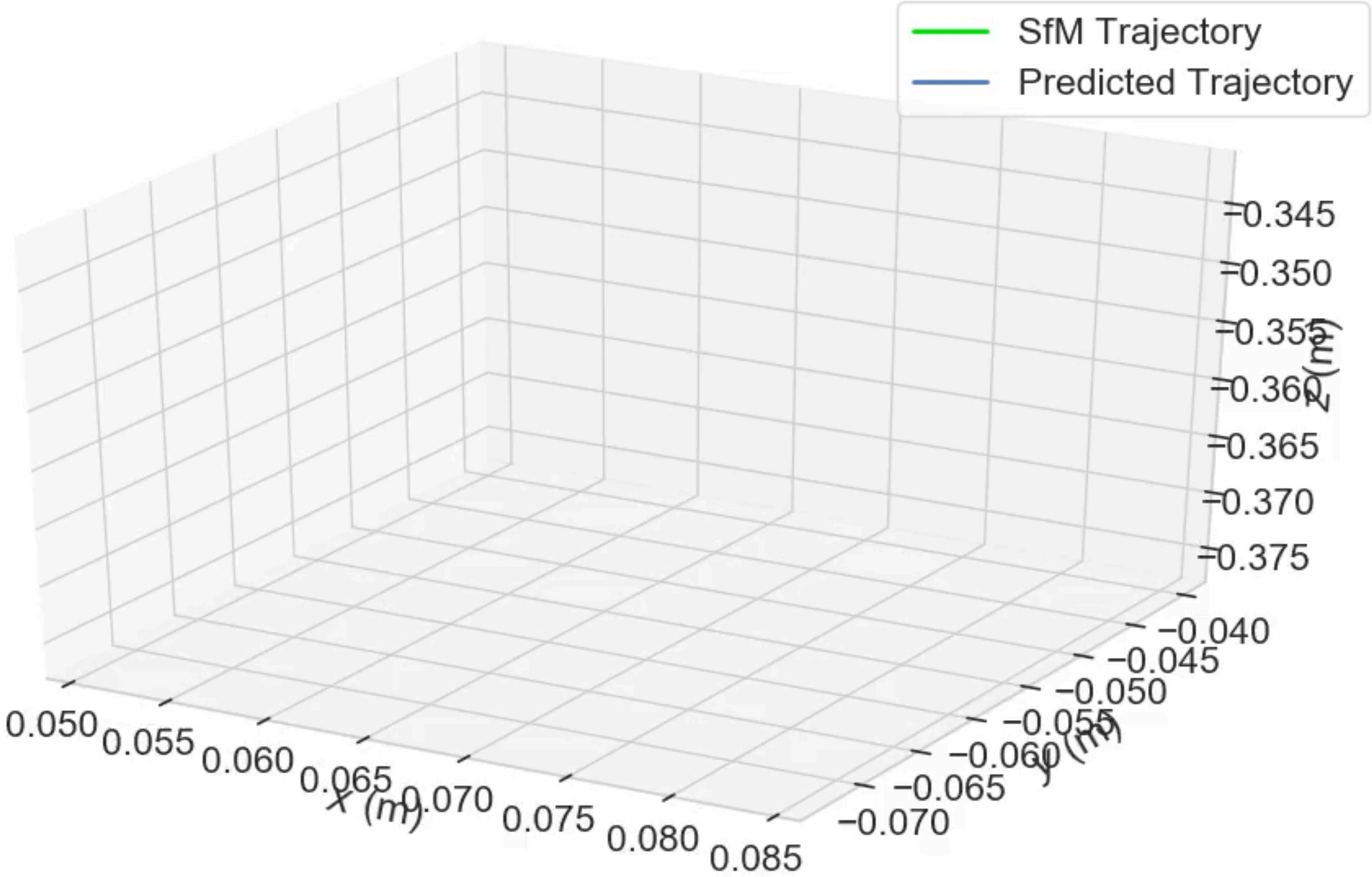
		Matterport 3D		MP3D → Replica	
Method	Pose	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Methods without any camera supervision					
Ours	×	20.58	0.64	21.72	0.77
Methods with full camera supervision					
Dosovisky et al.	√	14.79	0.57	14.36	0.68
Appearance Flow	√	15.87	0.53	17.42	0.66
SynSin (w/ voxel)	√	20.62	0.70	19.77	0.75
SynSin (w/ point cloud)	√	20.91	0.72	21.94	0.81

Results

Camera pose estimation

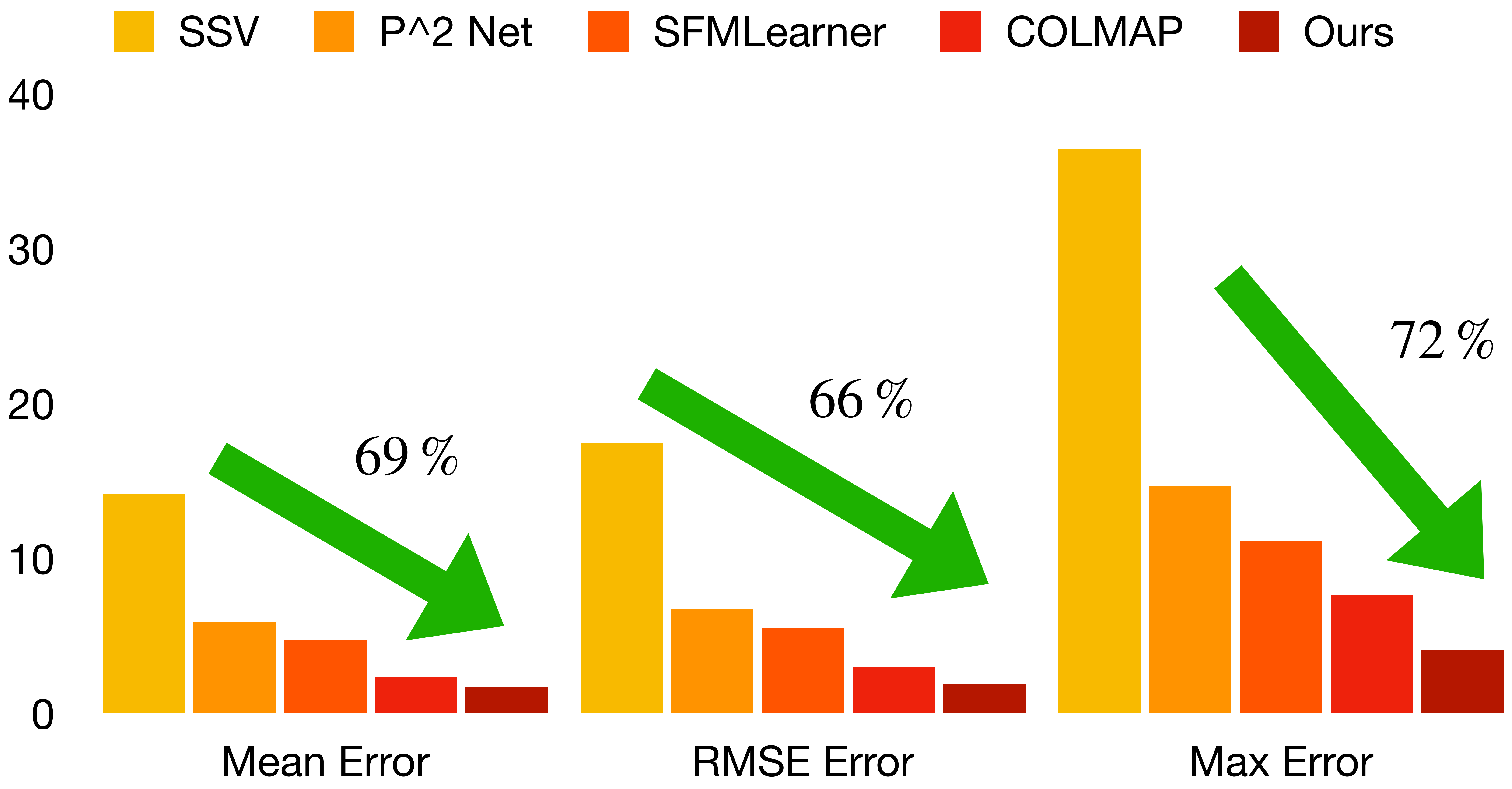
Input video

Trajectory Prediction



Results

Camera pose estimation



Comparisons on 30-frame videos

Results

Video following

Input image

Followed Video

Result

Camera Shaking

Results

Video following

Input image

Followed Video

Result

Rotating Right

Thanks!

Please see paper and website for details

https://zlai0.github.io/VideoAutoencoder