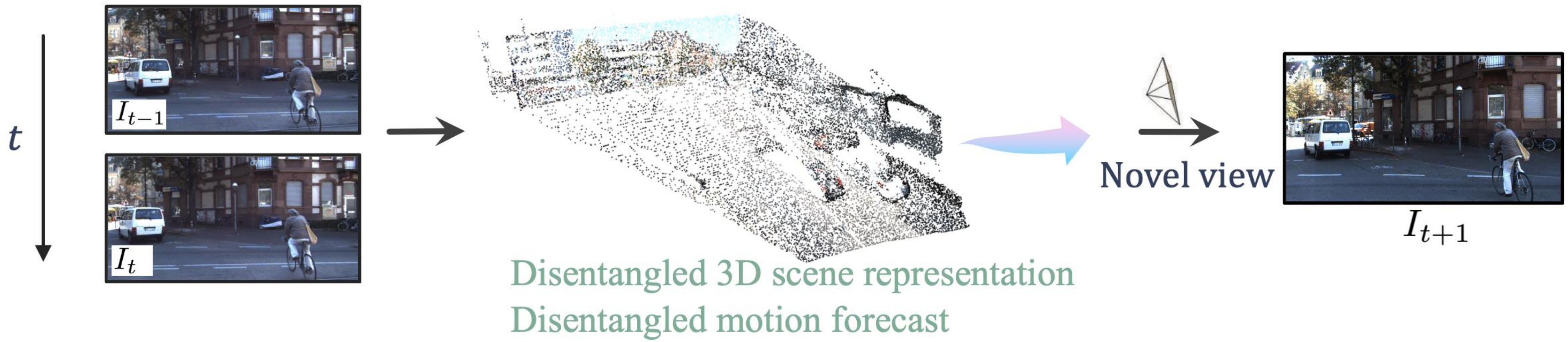
Paper & Code

Forecasting Future Videos from Novel Views via Disentangled 3D Scene Representation

Sudhir Yarram[1], Junsong Yuan[1]

[1]State University of New York at Buffalo

Motivation

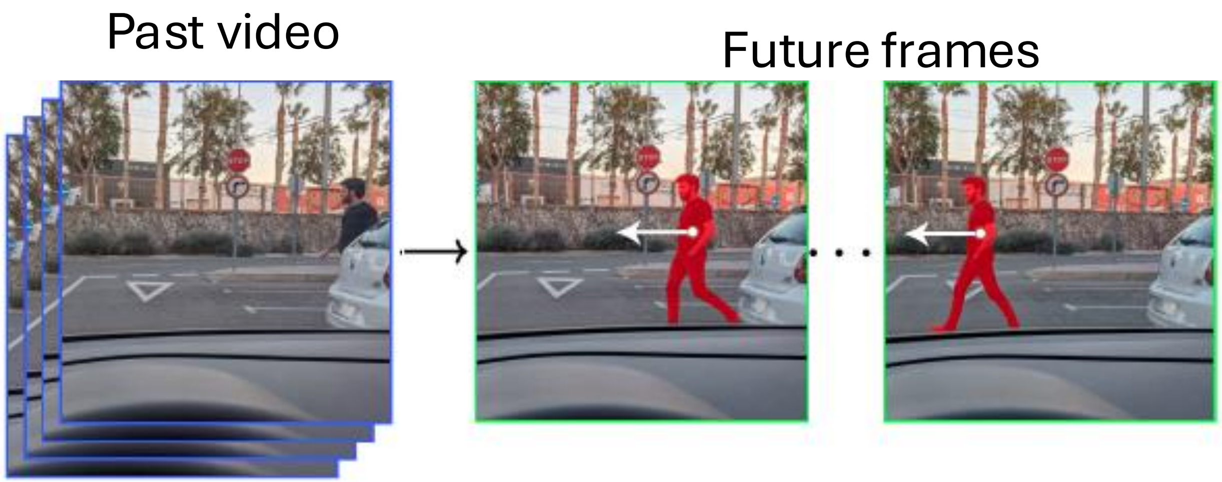


- The task of **Video extrapolation in space and time** (VEST) enables viewers to forecast a 3D scene into the future and view it from novel viewpoints.
- Our approach disentangles scene geometry from motion by lifting 2D scenes to 3D point clouds, enabling **high-quality** future video rendering from novel views.

Background

Future Forecasting:

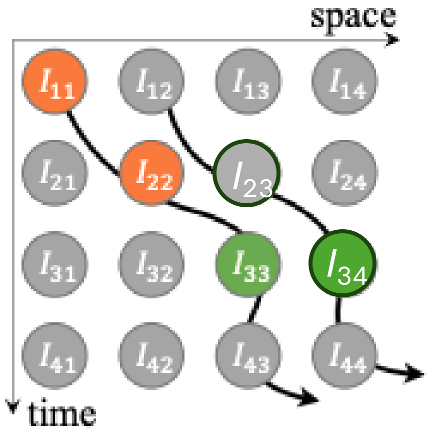
Video Prediction



Shi et. Al 2015

Future Forecasting

Video Prediction + Novel View Synthesis



Challenges

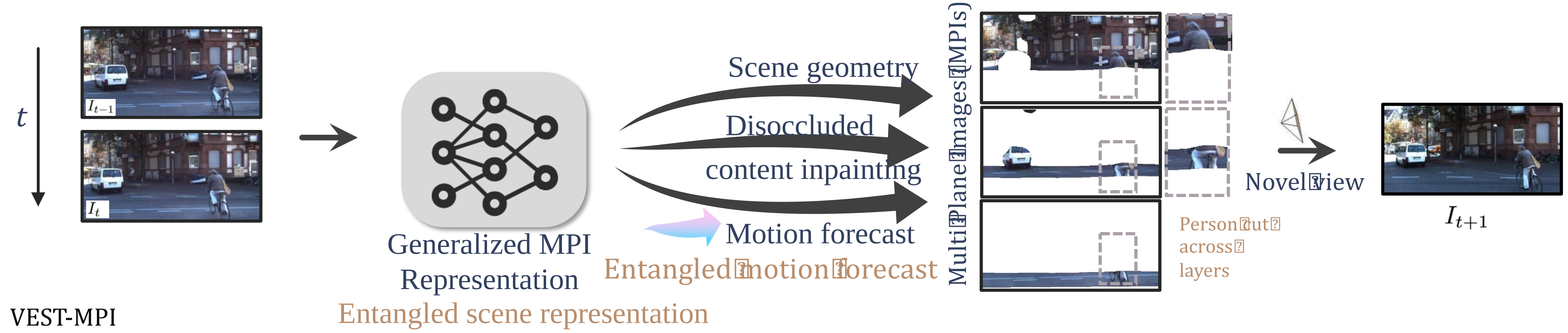
Three major challenges for future forecasting:

- ➢ Accurate estimation of scene geometry

- ➢ Forecasting future motion

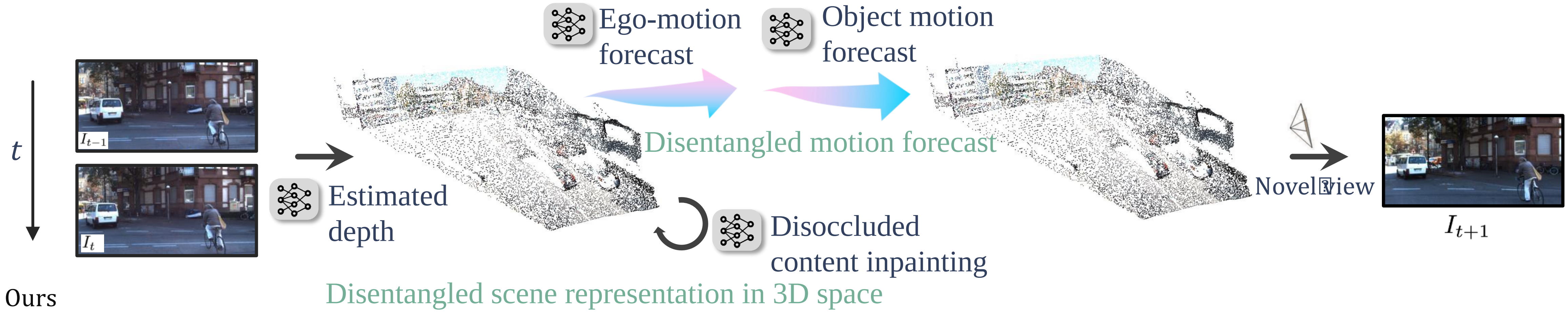
- ➢ Synthesizing disoccluded content

Related Work



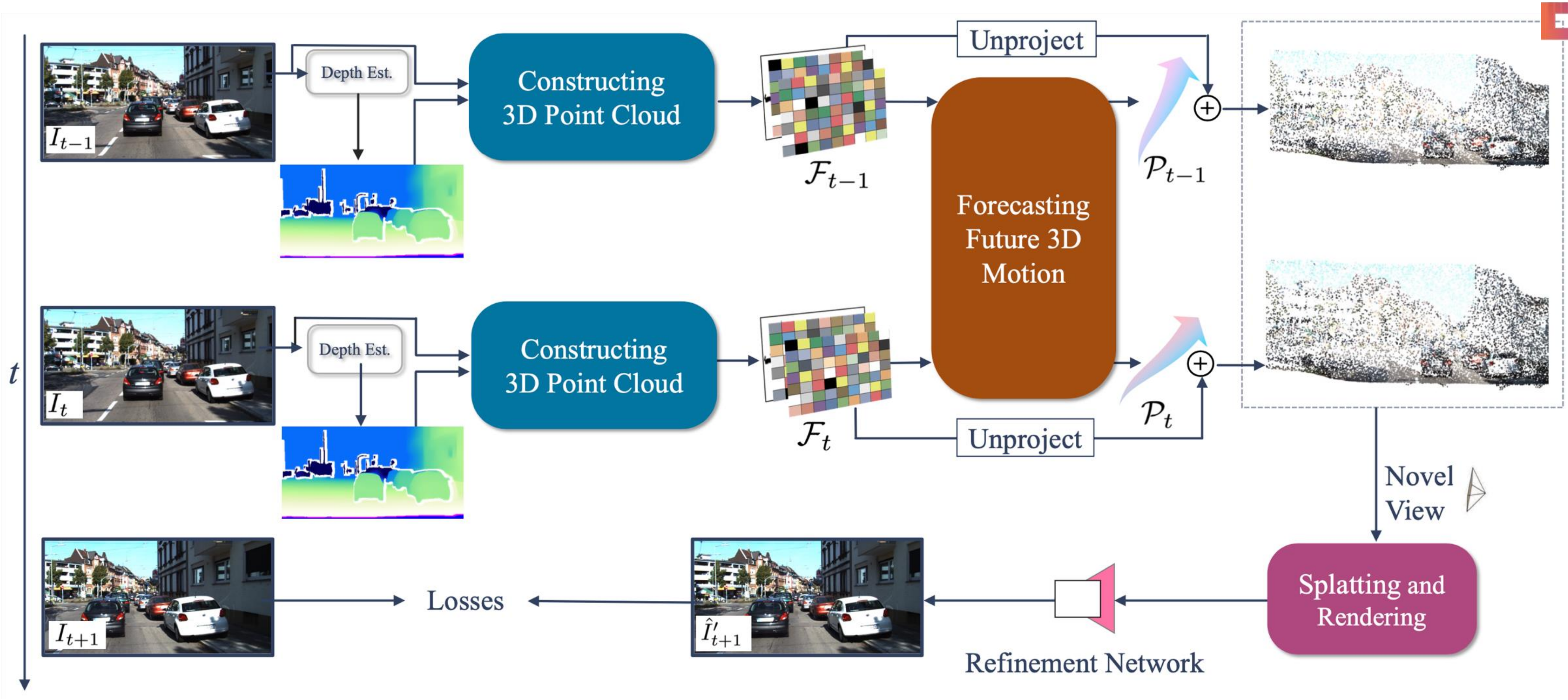
- Recent approaches propose
  - To learn an entangled representation
  - Aiming to model layered scene geometry, motion forecasting and novel view synthesis together
  - However, they rely on simplified affine motion and homography-based warping for each scene layer, resulting in inaccurate video extrapolation.

Approach



- Our approach
  - **Disentangles scene geometry** from scene motion, via lifting the 2D scene to 3D point clouds.
  - Additionally, we forecast **future 3D motion** by disentangling ego-motion of static objects from residual motion of dynamic objects.

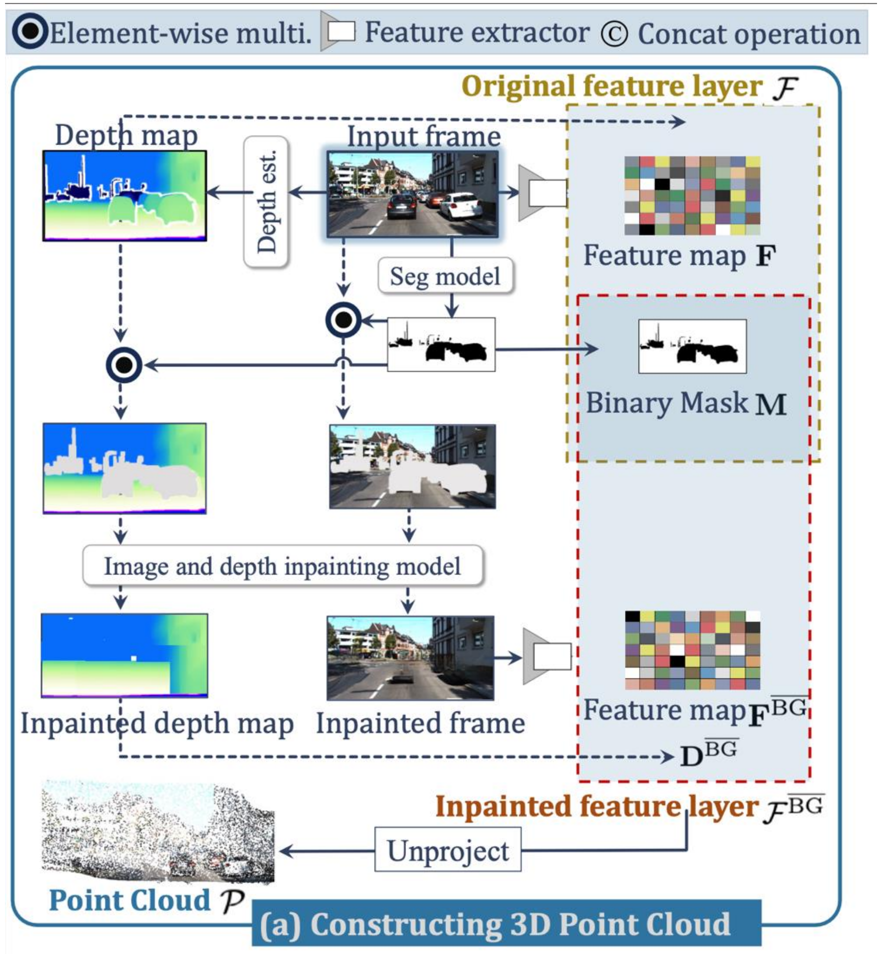
Approach

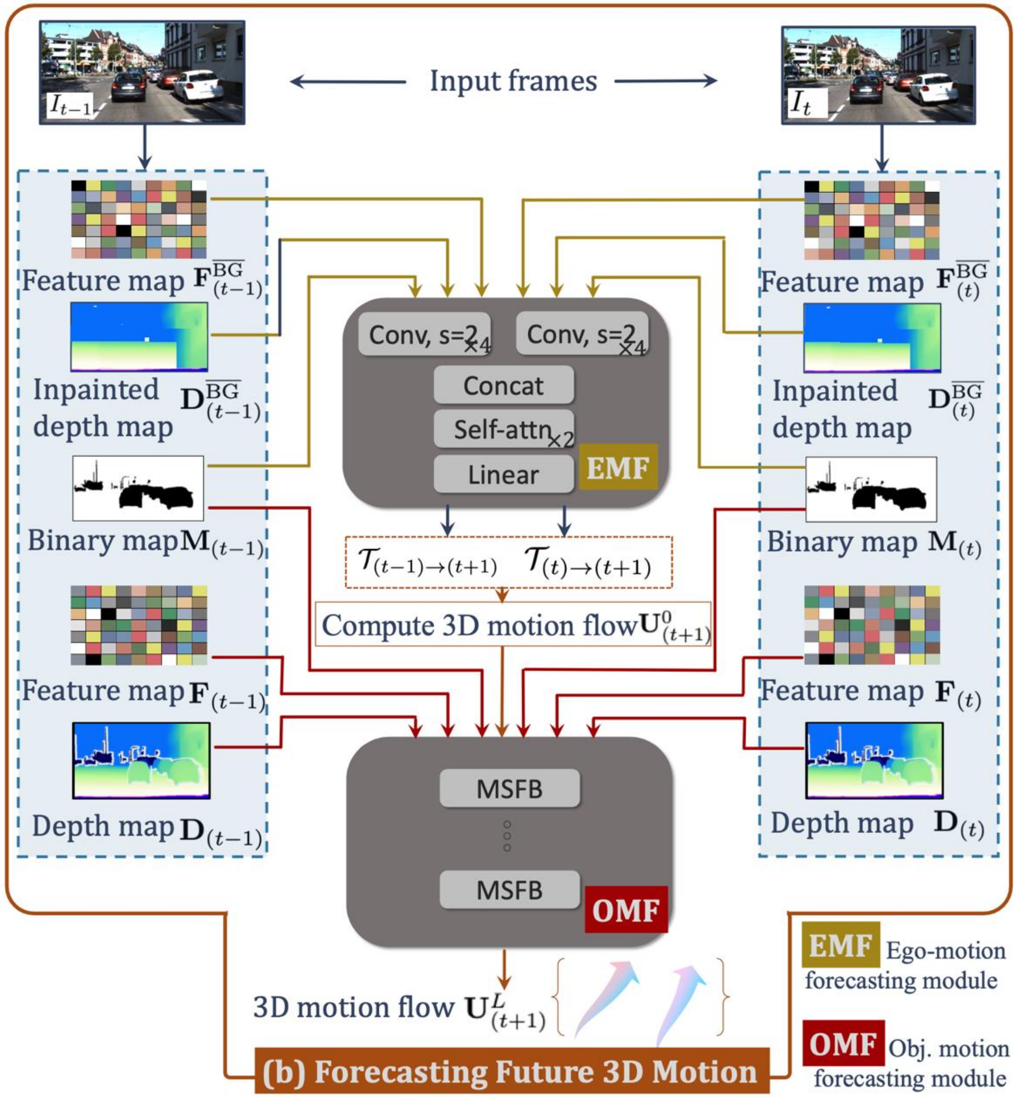


➢ Our framework aims to forecast a 3D scene into the future and view it from novel viewpoints. It comprises three primary steps:

- ❑ **Constructing 3D point clouds**
- ❑ **Forecasting future 3D motion**
- ❑ **Splatting and Rendering**

Approach





- For **Constructing 3D Point Cloud**, we leverage explicit 3D scene geometry (via depth estimation) to lift 2D scene into 3D point clouds.
- Forecasting Future 3D motion, results from both camera and object motions.
  - First, we forecast ego-motion by leveraging static background regions.
  - Then, predicted residual motion for dynamic objects (e.g., cars, persons).

Results

Quantitative results on Video Prediction Results on KITTI and Cityscapes

Method	Publication	Inputs	Cityscapes (512 × 1024)						KITTI (256 × 832)					
			$t+1$		$t+5$		$t+10$		$t+1$		$t+3$		$t+5$	
			SSIM↑	LPIPS↓	SSIM↑	LPIPS↓	SSIM↑	LPIPS↓	SSIM↑	LPIPS↓	SSIM↑	LPIPS↓	SSIM↑	LPIPS↓
CorrWise [10]	CVPR'22	R	92.8	8.5	83.9	15.0	75.1	21.7	82.0	17.2	73.0	22.0	66.7	25.9
SADM [1]	CVPR'21	R+L+F	95.9	7.6	83.5	14.9	N/A	N/A	83.1	14.4	72.4	24.6	64.7	31.2
DMVFN [13]	CVPR'23	R	95.7	5.6	83.5	14.9	N/A	N/A	**88.5**	10.7	78.0	19.3	70.5	26.0
WALDO [19]	ICCV'23	R+L+F	95.7	4.9	85.4	10.5	77.1	15.8	86.7	10.8	76.6	16.3	70.2	20.6
VEST-MPI [55]	ECCV'22	R	N/A	N/A	N/A	N/A	N/A	N/A	N/A	15.6	N/A	34.4	N/A	44.7
Ours		R+L+D	**96.4**	**4.6**	**86.2**	**9.8**	**78.0**	**14.9**	87.7	**10.1**	**77.6**	**15.4**	**71.3**	**19.8**

Quantitative results of Novel View Synthesis on KITTI

Method	Publication	In space only		In time only		
Extrapolation		LPIPS↓	SSIM↑	LPIPS $(\times 10^{-2})$↓		
				$t+1$	$t+3$	$t+5$
LDI [42]	ECCV'18	N/A	57.2		N/A	
MINE [20]	ICCV'21	10.8	82.2		N/A	
Tucker et al. [41]	CVPR'20	N/A	73.3		N/A	
PredRNNV2 [48]	TPAMI'22	N/A	N/A	30.8	45.7	54.2
VEST-MPI [55]	ECCV'22	8.5	82.5	11.5	28.8	39.1
Ours		**5.2**	**94.6**	**8.1**	**18.6**	**20.4**

Results

Qualitative Comparison with Baselines

Thank You!